Bayesian survival analysis in proportional hazard models with logistic relative risk

doi: 10.1111/j.1467-9469.2006.00543.x
 Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 34: 229–257, 2007

Bayesian Survival Analysis in Proportional Hazard Models with Logistic Relative Risk

PIERPAOLO DE BLASI

CEES, Department of Biology, University of Oslo, and Dipartimento di Statistica e Matematica Applicata, Università degli Studi di Torino

NILS LID HJORT

Department of Mathematics, University of Oslo

ABSTRACT. The traditional Cox proportional hazards regression model uses an exponential relative risk function. We argue that under various plausible scenarios, the relative risk part of the model should be bounded, suggesting also that the traditional model often might overdramatize the hazard rate assessment for individuals with unusual covariates. This motivates our working with proportional hazards models where the relative risk function takes a logistic form. We provide frequentist methods, based on the partial likelihood, and then go on to semiparametric Bayesian constructions. These involve a Beta process for the cumulative baseline hazard function and any prior with a density, for example that dictated by a Jeffreys-type argument, for the regression coefficients. The posterior is derived using machinery for Lévy processes, and a simulation recipe is devised for sampling from the posterior distribution of any quantity. Our methods are illustrated on real data. A Bernshteĭn–von Mises theorem is reached for our class of semiparametric priors, guaranteeing asymptotic normality of the posterior processes.

Key words: Bayesian semiparametrics, Bernshteĭn–von Mises theorem, Beta processes, hazard regression, Poisson random measures

1. Introduction and summary

We consider survival data with covariates, of the usual form (T_i, δ_i, x_i) for individuals $i = 1, \ldots, n$. Here, $T_i = \min(T_i^0, C_i)$ is the recorded lifetime, in terms of a life length T_i^0 that is observed if $\delta_i = 1$ but censored if $\delta_i = 0$, writing C_i for the censoring mechanism; also, x_i is a p-dimensional vector of covariates. The most popular methods for analysing such data remain those associated with Cox's proportional hazard rates model, which postulates that the hazard rates for the variables $T_1^0, \ldots, T_n^0$ take the form

$$\alpha_i(s) = \alpha(s) \exp(x_i^{\mathrm{t}} \beta) \quad \text{for } i = 1, \ldots, n, \tag{1}$$

with regression coefficients $\beta = (\beta_1, \ldots, \beta_p)$ accounting for the influence of covariates and with an unspecified baseline hazard rate function $\alpha(s)$ that represents the hazard for individuals with $x_i = 0$. See e.g. Andersen *et al.* (1993), hereafter ABGK, for extensive discussion of this model, including various detailed applications. The standard frequentist methods associated with (1) are based on the partial likelihood. Various Bayesian schemes have also been developed for handling this model; these need to be semiparametric in that one needs a nonparametric specification for the cumulative baseline hazard function $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ and a parametric prior for the β part (with or without prior independence between A and β). Among the more canonical Bayesian strategies is that of Hjort (1990b), which uses a Beta process prior for A coupled with any prior with a density for β.

This article is concerned with an important variation on (1), namely the model where the relative risk function takes a logistic form

$$\alpha_i(s)=\alpha(s)r(x_i^{\mathrm{t}}\gamma)=\alpha(s)\frac{\exp(x_i^{\mathrm{t}}\gamma)}{1+\exp(x_i^{\mathrm{t}}\gamma)} \quad \text{for } i=1,\ldots,n, \tag{2}$$

where $\gamma=(\gamma_1,\ldots,\gamma_p)$. The main point is to model situations where the relative risk function $r=r(w)$ is bounded, as opposed to the Cox model case where $r(w)=\exp(w)$ is unbounded. Thus if the relative risk is bounded by any constant, say $r(w)\le r_0$ for all w, then the r_0 may be subsumed into the $\alpha(s)$ part, so that, without loss of generality, we may work with r functions that take values in $[0,1]$. Of course other specific forms of r functions may be considered, in the same way as probit and other types of regression can be used instead of logistic regression for binomial data. We shall mainly stick to the specific version (2), however, but note that the frequentist and Bayesian methods we develop can be extended to other versions as well without severe efforts; see section 9. Aspects of the model (2) are discussed further in section 2, where we also provide arguments that make it plausible that the $r(w)$ function often will be bounded; that the model (2) may fit data better than the traditional model (1) is also demonstrated in section 7.

Frequentist analysis of the model (2) is not inherently much more difficult than that of the more familiar model (1), and uses as point of departure the partial likelihood function

$$L_n(\gamma)=\prod_{i=1}^{n}\left\{\frac{r(x_i^{\mathrm{t}}\gamma)}{nS_n^{(0)}(t_i,\gamma)}\right\}^{\delta_i},$$

in which

$$S_n^{(0)}(s,\gamma)=n^{-1}\sum_{i=1}^{n}Y_i(s)r(x_i^{\mathrm{t}}\gamma)$$

is the average relative risk at time s; here we write $Y_i(s)=I\{T_i\ge s\}$ for the still-at-risk indicator. Indeed, a paper by Prentice & Self (1983), following up on a classic paper of Andersen & Gill (1982) for the Cox model, gave results about the behaviour of the maxium partial likelihood estimator in models with a general relative risk function, and our favourite situation (2) is in essence covered by these general results. We indicate in section 3 that natural assumptions imply that the Prentice and Self type regularity conditions are in force, leading in particular to consistency and asymptotic normality;

$$\sqrt{n}(\hat\gamma-\gamma)\to_d N_p(0,\Sigma^{-1}), \tag{3}$$

where the Σ matrix in question may be estimated easily from data. Further details in this connection are in De Blasi (2006).

The main aim of the present article is, however, to develop natural *Bayesian* strategies for the model (2), and this is, in several respects, a harder task. This involves: (i) constructing a natural class of semiparametric priors, where care needs to be taken regarding the precise interpretation of (2), in that the most natural prior processes will not be continuous; (ii) deriving the precise nature of the posterior distribution of the model parameters (A,γ); (iii) developing workable recipes for Bayesian computation, e.g. via stochastic simulation from the posterior; and (iv) understanding the (frequentist) behaviour of Bayes estimators and of more general aspects of the posterior distribution.

Our semiparametric priors for (A,γ) involve Beta processes for A and e.g. Jeffreys-type priors for γ, and are discussed and developed in section 4. The somewhat laborious derivation of the resulting posterior distribution is given in section 5. A posterior simulation scheme is then developed in section 6. This is applied to a data set on Danish melanoma patient survival in section 7, where our general methods are illustrated to provide the distribution of the median remaining survival time for a patient with given covariates. In section 8, a so-called Bernshteĭn–von Mises theorem is provided, stating in essence that the Bayes machinery,

with our type of priors, leads to inferential statements that agree with the frequentist ones for large sample sizes; in particular, in parallel to the frequentist result (3), we prove that

$$\sqrt{n}(\gamma-\hat{\gamma})\,|\,\text{data} \;\to_d N_p(0,\Sigma^{-1}),$$

in probability. Such results are not to be taken for granted in Bayesian non- and semiparametrics, as examples have been exhibited in the literature that appear to have rather naturally constructed priors, but where the Bernshteĭn–von Mises theorem fails. When it holds, as it does for our semiparametric priors, it can be considered a stamp of approval for the Bayesian scheme. Section 9 ends with a list of concluding remarks, some pointing to further research questions. An appendix contains various technical lemmas and details required for proving our Bernshteĭn–von Mises theorem.

2. The logistic relative risk model

In this section, we show how various plausible background assumptions imply a bounded relative risk function, which is therefore an argument favouring model (2) over the traditional Cox model. We also briefly discuss other aspects of the model, including interpretation of its parameters.

2.1. Cumulative damage implies bounded relative risk

Instead of simply postulating the form of a model, and then perhaps confirm that a data set conforms to the assumed mathematical form (via goodness-of-fit analysis or model comparison methods), one may attempt to start the model building task at a deeper level, so to speak, taking as point of departure biologically plausible assumptions about the background processes associated with life lengths. By necessity, there are many such background scenarios worth investigating, and some of these lead to bounded relative risk functions, as we now demonstrate.

To keep matters relatively simple, consider a cumulative damage process of the compound Poisson type, say $Z(t)=\sum_{j\le\mathcal{M}(t)} G_j$ for $t\ge 0$, where $G_1, G_2, \ldots$ are independent and identically distributed (i.i.d.) shocks and $\mathcal{M}(t)$ is a Poisson process governing the frequency of these shocks, say with cumulative intensity function $\Lambda(t)=\int_0^t \lambda(s)\,\mathrm{d}s$. Imagine further that a person's survival function takes the form $S(t\,|\,\mathcal{H}_{t-})=\exp\{-Z(t)\}$ for $t\ge 0$, conditional on the damage history $\mathcal{H}_{t-}$ of what has taken place over the time interval $(0,t)$. The unconditional survival function then becomes

$$S(t)=E\exp\{-Z(t)\}=E\{E\exp(-G_1)\}^{\mathcal{M}(t)}=\exp\{-\rho\Lambda(t)\},$$

where $\rho=E\{1-\exp(-G_1)\}$. The point is that $\rho\le 1$, regardless of the distribution of the shocks. Thus the life-length distribution must have hazard rate $\rho\lambda(s)$ with ρ bounded by 1.

Various special cases may be investigated, with different assumptions leading to different hazard regression models. If different individuals experience shocks with about the same regularity, corresponding to the same $\lambda(s)$ intensity, but with the sizes of shocks influenced by covariates x, then the implied model is $\alpha(s,x)=\rho(x)\lambda(s)$, with $\rho(x)\in(0,1)$. This is exactly what happens for our model (2). Other variations on this theme, some lending support to the bounded relative risk function assumption and others to perhaps intermediate models, are discussed in Aalen & Hjort (2002) and Hjort (2003). The main point we are making is that fully plausible background assumptions about what causes a life to end imply a bounded relative risk function, at odds with the most traditional Cox model (see also Gjessing *et al.*, 2003, and section 9 below).

2.2. Interpretation of parameters

For the logistic form of $r(\cdot)$, how the covariates x are scaled does matter crucially for the interpretation of both the baseline hazard $\alpha(\cdot)$ and the regression coefficients $\gamma_1, \ldots, \gamma_p$. An individual with covariate $x=0$ has hazard rate $(1/2)\alpha(s)$. We recommend that each covariate scale is centred, to have mean value (or perhaps median value) equal to zero. This helps interpretation of $\alpha(s)$, as twice the hazard rate of an 'average individual' whose covariates are all zero, as well as of the γ_j coefficients. Having covariates on both sides of zero also makes the partial likelihood more peaked and helps stabilizing the numerical procedures that we develop in later sections, associated with the Bayesian semiparametric strategies.

The relative risk between two individuals may be unbounded, even with the bounded form of $r(w)$; two individuals with covariates x_1 and x_2 have hazard ratio $r(x_1^{\mathrm{t}}\gamma)/r(x_2^{\mathrm{t}}\gamma)$, which has unlimited range (as for the Cox model). Note that the traditional exponential form of (1) has the risk of overdramatizing hazard assessment for individuals with extreme covariates; the logistic model is safer in this regard. We also point out that the model (2) should not be fitted with an intercept term; we employ only real covariates in the $r(x_i^{\mathrm{t}}\gamma)$ term. A model with hazards taking the form $\alpha(s)r(\gamma_0 + x_i^{\mathrm{t}}\gamma)$ becomes overdetermined. For further detailed discussion of interpretation issues, and of differences between the Cox model and the new logistic form model, see De Blasi (2006).

3. Partial likelihood and frequentist inference

Before we start climbing our Bayesian mountain, we briefly describe how matters are handled in the frequentist camp. The main output of the analysis is the maximum partial likelihood estimator $\hat{\gamma}$ and its associated estimated covariance matrix (which then leads to e.g. confidence intervals for the γ_j parameters of (2)). Showing consistency and asymptotic normality is accomplished following either Prentice & Self (1983) or, for a more streamlined argument that also requires less restrictive regularity conditions, Hjort & Pollard (1993). Investigating these matters requires our working with certain counting processes, at-risk processes and martingales, quantities that we shall also need later when we explore the behaviour of Bayes estimators. In this section, we postulate that model (2) is in force with a parameter vector γ_{tr} and for a certain positive $\alpha(\cdot)=\alpha_{\mathrm{tr}}(\cdot)$ function, with cumulative $A_{\mathrm{tr}}(t)=\int_0^t \alpha_{\mathrm{tr}}(s)\,\mathrm{d}s$, and that observations are recorded over a fixed and finite time window $[0,\tau]$. Thus $(A_{\mathrm{tr}}, \gamma_{\mathrm{tr}})$ determines the true model, under which our large-sample results are derived.

In addition to the at-risk indicator $Y_i(s)$, we introduce the counting process $N_i(t)=I\{T_i \le t, \delta_i = 1\}$ that jumps precisely at time point T_i, provided this observation is non-censored, and the martingale

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) r(x_i^{\mathrm{t}}\gamma_{\mathrm{tr}})\,\mathrm{d}A_{\mathrm{tr}}(s) \quad \text{for } t \ge 0.$$

We shall also have occasion to use the accumulated counting process $N_\cdot = \sum_{i=1}^n N_i$. For our model (2) the log-partial likelihood function can be written

$$\ell_n(\gamma) = \sum_{i=1}^n \int_0^\tau \left[\log r(x_i^{\mathrm{t}}\gamma) - \log\{nS_n^{(0)}(s,\gamma)\}\right] \mathrm{d}N_i(s), \tag{4}$$

where $r(\cdot)$ has the logistic form. A full analysis of all details pertaining to the large-sample behaviour of $\hat{\gamma}$ requires working with r' and r'', the derivatives of $r(w)$, and with u' and u'', the derivatives of $u(w)=\log r(w)$. For the logistic model, $r'=r(1-r)$, $r''=r(1-r)(1-2r)$, $u'=1-r$ and $u''=-r(1-r)$. It is also useful to work with

$$S_n^{(1)}(s,\gamma)=n^{-1}\sum_{i=1}^{n} Y_i(s)r'(x_i^{\mathrm{t}}\gamma)x_i,$$

which is the derivative of $S_n^{(0)}(s,\gamma)$ w.r.t. γ, and

$$S_n^{(2)}(s,\gamma)=n^{-1}\sum_{i=1}^{n} Y_i(s)u'(x_i^{\mathrm{t}}\gamma)^2 r(x_i^{\mathrm{t}}\gamma)x_i x_i^{\mathrm{t}}.$$

Furthermore, let $E_n(s,\gamma)=S_n^{(1)}(s,\gamma)/S_n^{(0)}(s,\gamma)$ and

$$\begin{aligned} V_n(s,\gamma)&=\frac{S_n^{(2)}(s,\gamma)}{S_n^{(0)}(s,\gamma)}-E_n(s,\gamma)E_n(s,\gamma)^{\mathrm{t}}\\ &=n^{-1}\sum_{i=1}^{n}\frac{Y_i(s)r(x_i^{\mathrm{t}}\gamma)}{S_n^{(0)}(s,\gamma)}\{u'(x_i^{\mathrm{t}}\gamma)x_i-E_n(s,\gamma)\}\{u'(x_i^{\mathrm{t}}\gamma)x_i-E_n(s,\gamma)\}^{\mathrm{t}}. \end{aligned}$$

The essential requirements that secure consistency and asymptotic normality are that the average processes defined above converge to suitable limit functions, say $s^{(0)}(s,\gamma)$, $s^{(1)}(s,\gamma)$, $s^{(2)}(s,\gamma)$, uniformly in $s\in[0,\tau]$ and over a neighbourhood of γ_{tr}. This also implies corresponding convergence of $E_n(s,\gamma)$ and $V_n(s,\gamma)$ to limit functions $e(s,\gamma)$ and $v(s,\gamma)$. In particular, there is convergence in probability

$$\Sigma_n=n^{-1}\sum_{i=1}^{n}\int_0^{\tau} V_n(s,\gamma_{\mathrm{tr}})\,\mathrm{d}N_i(s)\to_p\Sigma=\int_0^{\tau} v(s,\gamma_{\mathrm{tr}})s^{(0)}(s,\gamma_{\mathrm{tr}})\,\mathrm{d}A_{\mathrm{tr}}(s).$$

One may now prove that $\sqrt{n}(\hat{\gamma}-\gamma_{\mathrm{tr}})\to_d N_p(0,\Sigma^{-1})$, under mild regularity conditions. In addition, a consistent estimator for the limit covariance matrix is

$$\hat{\Sigma}=n^{-1}\sum_{i=1}^{n} V_n(t_i,\hat{\gamma})\delta_i.$$

A strong assumption that secures these convergence statements is that the sequence of triples (T_i^0,C_i,x_i) is i.i.d. with a finite fourth moment for the covariate distribution (see Prentice & Self, 1983). Also, De Blasi (2006) provides a detailed discussion, showing that a small set of regularity conditions in this framework imply the regularity conditions A–F in Prentice & Self (1983). These assumptions can be substantially weakened, though, without losing the conclusions summarized above, if one follows the line of arguments given in Hjort & Pollard (1993) for the ordinary Cox model.

For the sake of completeness we also state, without proof, the asymptotic normality of the Breslow–Aalen type estimator

$$\hat{A}(t)=\int_0^{t} n^{-1}\sum_{i=1}^{n}\frac{\mathrm{d}N_i(s)}{S_n^{(0)}(s,\hat{\gamma})}.$$

The result is that $\sqrt{n}(\hat{\gamma}-\gamma_{\mathrm{tr}})$ and the stochastic process

$$\sqrt{n}\{\hat{A}(t)-A_{\mathrm{tr}}(t)\}+\sqrt{n}(\hat{\gamma}-\gamma_{\mathrm{tr}})^{\mathrm{t}}\int_0^{t}\mathrm{e}(s,\gamma_{\mathrm{tr}})\alpha_{\mathrm{tr}}(s)\,\mathrm{d}s \tag{5}$$

are asymptotically independent, the latter converging in distribution to a Gaussian martingale with variance function $\int_0^t s^{(0)}(s,\gamma_{\mathrm{tr}})^{-1}\alpha_{\mathrm{tr}}(s)\,\mathrm{d}s$. Proving (5) is again accomplished applying the theory laid out in Prentice & Self (1983) or Hjort & Pollard (1993).

4. The Beta process and the Jeffreys prior

This section describes a natural class of prior distributions for (A, γ). These are placed on infinite-dimensional parameter spaces, so care is required in their specification. There is also a crucial difference between the Cox model (1) and the logistic model (2) which matters rather more in the Bayesian formulations than for the frequentist treatment we gave in the previous section. This is related to the problem of deciding on a good version of the model that also allows jumps in the cumulative hazard processes.

4.1. Bayesian approaches for the Cox model

Before we start working with the logistic model we briefly review approaches that have been taken to handle the Cox model from the Bayesian point of view. Kalbfleisch (1978) used Doksum's (1974) neutral to the right processes; these are random cumulative distribution functions of the form $F(t)=1-\exp\{-Z(t)\}$ with Z having independent non-negative increments, i.e. being a Lévy process. This was later generalized by Wild & Kalbfleisch (1981), extending earlier work of Ferguson & Phadia (1979) from settings of i.i.d. data to situations with covariates. Hjort (1990b) chose to work directly in the space of cumulative hazard functions $\mathcal{A}$, as opposed to using $-\log\{1-F(t)\}$ type constructions. Formally, $\mathcal{A}$ is the set of right-continuous non-decreasing functions A on the halfline, starting with value zero at zero, such that each jump $\Delta A(t)=A(t)-A(t-)$ is in $[0, 1]$. This constraint is crucial for the hazard interpretation to be valid,

$$\mathrm{d}A(s)=\Pr\{\text{die in}\,[s, s+\mathrm{d}s]\,|\,\text{survived to time } s\}. \tag{6}$$

The *Beta processes* (Hjort, 1990b) have paths that with probability one are in $\mathcal{A}$, and have become the perhaps most popular type of priors for use in models with hazard rates. A Beta process A with parameters $(c(\cdot), A_0(\cdot))$, where A_0 is a fixed cumulative hazard rate and c a positive concentration function, has independent increments with the property that

$$\mathrm{d}A(s)\approx_d \mathrm{Beta}[c(s)\,\mathrm{d}A_0(s), c(s)\{1-\mathrm{d}A_0(s)\}], \tag{7}$$

infinitesimally speaking. This gives the perhaps best intuitive understanding of a Beta process; for example, note that $\mathrm{d}A(s)$ has mean $\mathrm{d}A_0(s)$ and variance $\mathrm{d}A_0(s)\{1-\mathrm{d}A_0(s)\}/\{1+c(s)\}$, indicating that A_0 is the prior guess for the cumulative hazard function while c acts as a concentration function; bigger c makes for tighter concentration around A_0, and vice versa. More formalistic definitions of Beta processes are required below, when we wish to prove results about the posterior distribution.

To work with Beta processes for the Cox model one needs to avoid a specification of that model where hazard jumps can be bigger than one. A natural solution, worked with in Hjort (1990b), is the framework where individuals have cumulative hazard rate functions $A_1, \ldots, A_n$, and where

$$1-\mathrm{d}A_i(s)=\{1-\mathrm{d}A(s)\}^{r(x_i^{\dagger}\beta)} \quad \text{for } i=1,\ldots,n. \tag{8}$$

This implies, via the product integral transformation, that $S_i(t)=\pi_{[0,t]}\{1-\mathrm{d}A_i(s)\}$ is equal to $S(t)^{r(x_i^{\dagger}\beta)}$, where S is the survival curve associated with cumulative hazard rate A. A quite natural class of priors is now to let A be a Beta process and β have any prior density $\pi(\beta)$. The posterior distribution for (A, β), along with various properties, was worked out in Hjort (1990b).

Later references, regarding Bayesian handling of the Cox model, include Laud *et al.* (1998), who investigated numerical implementations of Hjort's methods, and Kim & Lee (2003a, 2003b), who gave connections from Hjort's methods to Bayesian bootstraps for survival data.

Yet further options are worked by Clayton (1991) via Gamma processes, Rigat (2004) using Beta–Stacy priors, Walker & Mallick (1997) with Pólya trees, Gelfand & Kottas (2003) for median survival time regression via Dirichlet processes and Ishwaran & James (2004), employing kernel mixtures of weighted Gamma process priors for the baseline cumulative hazard. Finally, Hjort & Kim (2007) investigate different constructions that involve mixtures of Beta processes.

4.2. The Beta process prior for A

For the logistic relative risk model one option is to follow (8), with the logistic form for $r(x_i^{\mathrm{t}}\gamma)$. This would proceed in a manner paralleling the methods and results alluded to above. In this article, we choose instead to stay with the framework

$$\mathrm{d}A_i(s) = \mathrm{d}A(s) r(x_i^{\mathrm{t}}\gamma) \quad \text{for } i = 1, \ldots, n, \tag{9}$$

making the hazard functions (and their cumulatives) directly proportional to each other. This is a viable approach for the logistic form of the model, as $r(w) \in (0, 1)$, but not for the Cox model, as jumps then would risk slipping outside $(0, 1)$, which clashes with the fundamental hazard interpretation (6).

The class of priors we shall work with takes therefore a Beta process $(c(\cdot), A_0(\cdot))$ for A and any prior with a density for γ, say $\pi(\gamma)$. As mentioned above we need more formally precise definitions of Beta processes, in order to accurately derive the description of the posterior. There are (at least) two useful ways of doing this.

The first is via *Lévy representations*. Let $A_0 \in \mathcal{A}$ with jumps at points $\{t_1, t_2, \ldots\}$ and let $c(\cdot)$ be a piecewise continuous, non-negative real-valued function on $[0, \infty)$. Then the Beta process (c, A_0) admits the representation

$$E \exp\{-\theta A(t)\} = \left[\prod_{j: t_j \le t} E \exp\{-\theta \Delta A(t_j)\}\right] \exp\left\{-\int_0^1 (1 - \mathrm{e}^{-\theta s})\, \mathrm{d}L_t(s)\right\} \tag{10}$$

for $t \ge 0$, where $\Delta A(t_j)$ is the jump size at position t_j, with Beta distribution

$$\Delta A(t_j) \sim \text{Beta}\,[c(t_j)\Delta A_0(t_j), c(t_j)\{1 - \Delta A_0(t_j)\}]. \tag{11}$$

The continuous Lévy measure takes the form

$$\mathrm{d}L_t(s) = \int_0^t s^{-1}(1 - s)^{c(z)-1} c(z)\, \mathrm{d}A_{0,c}(z)\, \mathrm{d}s \tag{12}$$

for $t \ge 0$, where $A_{0,c}(t) = A_0(t) - \sum_{t_j \le t} \Delta A_0(t_j)$.

The second mathematical path to follow is that of *Poisson random measures*, as in Kim (1999) and Hjort & Kim (2007); the following definition is made wide enough to encompass the posterior distributions we later need to describe. For any given Lévy process A with paths in $\mathcal{A}$, there is a unique random measure μ on $[0, 1] \times [0, \infty)$, such that $\mu(\mathrm{d}s, \mathrm{d}z) = I\{\Delta A(z) \in [s, s + \mathrm{d}s]\}$, with correspondence formula, from μ to A, of the form $A(t) = \int_0^t \int_0^1 s\mu(\mathrm{d}s, \mathrm{d}z)$. The Poisson random measure μ has a compensator or mean function ν, i.e. a sigma-finite measure on $[0, 1] \times \mathcal{R}_+$, such that $\int_0^t \int_0^1 s\nu(\mathrm{d}s, \mathrm{d}z)$ is finite for all $t \ge 0$. The connection with the Lévy representation in (10) is that ν extends the continuous measure $\mathrm{d}L_t(s)$ by incorporating the distribution of the fixed points of discontinuity.

A Beta process $A \sim \text{Beta}(c(\cdot), A_0(\cdot))$ with continuous A_0 is characterized by

$$\nu(\mathrm{d}s, \mathrm{d}t) = a(s, t)\, \mathrm{d}s\, \mathrm{d}A_0(t), \quad \text{where } a(s, t) = s^{-1}(1 - s)^{c(t)-1} c(t). \tag{13}$$

In case of fixed discontinuities at $t_1, \ldots, t_m$, (13) becomes

$$v(\mathrm{d}s, \mathrm{d}t) = a(s,t)\,\mathrm{d}s\,\mathrm{d}A_{0,c}(t) + \sum_{j=1}^{m} f_{t_j}(s)\,\mathrm{d}s\ \delta_{t_j}(\mathrm{d}t), \tag{14}$$

where $f_{t_j}(s)$ is the Beta density of the jump at t_j, see equation (11), whereas δ_{t_j} is the Dirac measure with unit mass at point t_j. The expectation of $A(t)$ is recovered as

$$EA(t) = \int_0^t \int_0^1 sv(\mathrm{d}s, \mathrm{d}z), \tag{15}$$

which by (14) splits into the mean of a continuous part and $\sum_{t_j \le t} \xi_j$, where ξ_j is the mean of the jump distribution f_{t_j} at t_j.

4.3. The Jeffreys prior for γ

Specifying a meaningful prior distribution for the regression parameters is in general a difficult task. It can be particularly difficult in multivariate cases or when little prior information is available. In such cases reference priors can be used. Here we derive the formula for the Jeffreys prior, which, for general parametric models, is proportional to $|J_n(\gamma)|^{1/2}$, the square root of the determinant of the information matrix. This formula is the solution to the natural question of invariance of prior belief statements under arbitrary smooth transformations of the parameter vector (see e.g. Gelman *et al.*, 2004, Ch. 2).

To derive the Jeffreys prior we start with the log-likelihood for the case of a known baseline hazard $\alpha(s)$ and assume no censoring. This is reasonable since the form of the prior for γ should not depend on the amount of censoring or on whether $\alpha(\cdot)$ is known or not. In this setting, the log-likelihood function $\log L_n(\gamma)$ is

$$\sum_{i=1}^{n} \int_0^{\tau} \{\log \alpha(s) r(x_i^{\mathrm{t}}\gamma)\,\mathrm{d}N_i(s) - Y_i(s)\alpha(s) r(x_i^{\mathrm{t}}\gamma)\,\mathrm{d}s\},$$

(see e.g. ABGK, Ch. VI). This leads to

$$\begin{aligned}\frac{\partial^2 \log L_n(\gamma)}{\partial\gamma\partial\gamma^{\mathrm{t}}} &= \sum_{i=1}^{n} \int_0^{\tau} \{u''(x_i^{\mathrm{t}}\gamma)\,\mathrm{d}N_i(s) - Y_i(s)\alpha(s) r''(x_i^{\mathrm{t}}\gamma)\,\mathrm{d}s\} x_i x_i^{\mathrm{t}} \\ &= \sum_{i=1}^{n} \{u''(x_i^{\mathrm{t}}\gamma) - A(T_i^0) r''(x_i^{\mathrm{t}}\gamma)\} x_i x_i^{\mathrm{t}}.\end{aligned}$$

Recall that T_i^0 has cumulative hazard rate $A(t)r(x_i^{\mathrm{t}}\gamma)$, which implies that $A(T_i^0)$ has mean $1/r(x_i^{\mathrm{t}}\gamma)$. Hence the Fisher information matrix becomes

$$J_n(\gamma) = \sum_{i=1}^{n} \left\{ -u''(x_i^{\mathrm{t}}\gamma) + \frac{r''(x_i^{\mathrm{t}}\gamma)}{r(x_i^{\mathrm{t}}\gamma)} \right\} x_i x_i^{\mathrm{t}} = \sum_{i=1}^{n} \left\{ \frac{r'(x_i^{\mathrm{t}}\gamma)}{r(x_i^{\mathrm{t}}\gamma)} \right\}^2 x_i x_i^{\mathrm{t}}.$$

The Jeffreys recipe thus gives a flat prior for the Cox model (as found in Hjort, 1986), and

$$\pi(\gamma) \propto \left| n^{-1} \sum_{i=1}^{n} \frac{x_i x_i^{\mathrm{t}}}{\{1 + \exp(x_i^{\mathrm{t}}\gamma)\}^2} \right|^{1/2} \tag{16}$$

for the logistic relative risk hazard regression model (2). The prior (16) for γ is bounded, and improper in the sense of having infinite integral.

In section 5, we show how the marginal posterior of γ is proportional to the prior times a term that corresponds to the likelihood after integrating out the conditional distribution of A given γ and data. It follows that the (16) prior leads to a proper posterior distribution

as long as the integrated likelihood has finite integral, as the Jeffreys prior stays bounded in probability if only the second moment of the covariance distribution is finite, i.e. $\pi(\gamma) = O_p(1)$. In appendix A, we show that the integrated likelihood can essentially be bounded from above by a bounded factor times the partial likelihood. Hence the problem translates to the tails of the partial likelihood and the condition of covariates centred around zero roughly implies its correct behaviour. Alternatively, arguments as in the proof of theorem 4.1 in Kim & Lee (2003a) can be used here, to show that the posterior is proper, typically for all $n > p$.

5. The posterior distribution

Section 4 described our class of priors, with a Beta process for A and any prior with a density for γ, like the Jeffreys prior (16). In this section, we derive the explicit form of the posterior distribution, given a set of survival data with covariates. This task is not a simple Bayes theorem exercise, as the probability distributions are intricate constructions on infinite-dimensional spaces, and with both discrete and continuous components. The task is split into two: we first find the posterior distribution of A given γ, and then the marginal posterior of γ. In section 6, we describe a Gibbs sampling scheme for drawing samples from the joint posterior distribution of (A, γ), and illustrate this in section 7 for a case of median remaining lifetime.

5.1. The posterior distribution of A given γ

For a sample of right censored data $(t_1, x_1, \delta_1), \ldots, (t_n, x_n, \delta_n)$, let $t_1 < \cdots < t_{q_n}$ denote the complete set of distinct observations. We do assume that the censoring mechanism has worked independently of the distribution of real lifetimes. The joint likelihood can be expressed by means of a product integral of binomials,

$$L_n(A, \gamma) = \prod_{i=1}^{n} \prod_{[0,\tau]} \mathrm{d}A(s, x_i)^{\mathrm{d}N_i(s)} \{1 - \mathrm{d}A(s, x_i)\}^{Y_i(s) - \mathrm{d}N_i(s)}, \tag{17}$$

where $\mathrm{d}A(s, x_i) = \mathrm{d}A(s) r(x_i^{\mathrm{t}}\gamma)$ by (9). For a discussion and interpretation of (17) see ABGK, section IV.1.5. Here we find

$$\begin{aligned} L_n(A, \gamma) = &\prod_{i=1}^{q_n} \left[\prod_{j=1}^{n} \{r(x_i^{\mathrm{t}}\gamma) \Delta A(t_i)\}^{\Delta N_j(t_i)} \{1 - r(x_j^{\mathrm{t}}\gamma) \Delta A(t_i)\}^{Y_j(t_i) - \Delta N_j(t_i)} \right] \\ &\times \prod_{\mathrm{out}} \prod_{i=1}^{n} \{1 - r(x_i^{\mathrm{t}}\gamma)\, \mathrm{d}A(t)\}^{Y_i(t)}, \end{aligned} \tag{18}$$

where 'out' is the union of all time intervals where nothing has happened, i.e. between observed lifetimes; $\mathrm{out} = \{t \in [0, \tau] : \Delta N_\cdot(t) = 0\}$. The key point now is that when the prior of A is a Beta process, the posterior will not be quite as simple, but it remains a Lévy process with Lévy measure concentrated on $[0, 1]$.

The following lemma is a necessary building block for the results to follow; cf. also lemma A.3 in Hjort (1990b). Let $\mathcal{M} = \{t_1, \ldots, t_m\}$ denote the set of discontinuity points for A_0 and let $A_{0,c}(t) = A_0(t) - \sum_{t_j \in \mathcal{M} : t_j \le t} \Delta A_0(t_j)$.

Lemma 1

Let A be a Beta process (c, A_0), where A_0 has jumps in $\mathcal{M}$, let r be a number in $[0, 1]$, and let $\theta(\cdot)$ be any piecewise constant function. Then the mean value of $\prod_{[0,t]} \{1 - r\, \mathrm{d}A(z)\} \times \exp\{-\int_0^\infty \theta(z)\, \mathrm{d}A(z)\}$ may be expressed as

$$\prod_{t_j \le t} E\{1 - r\Delta A(t_j)\} \exp\{-\theta(t_j)\Delta A(t_j)\} \times \prod_{t_j > t} E \exp\{-\theta(t_j)\Delta A(t_j)\}$$
$$\times \exp\left\{ - \int_0^t \int_0^1 \{1 - \mathrm{e}^{-\theta(z)s} + rs\,\mathrm{e}^{-\theta(z)s}\} a(s,z)\,\mathrm{d}s\,\mathrm{d}A_{0,c}(z) \right\}$$
$$\times \exp\left[- \int_t^\infty \int_0^1 \{1 - \mathrm{e}^{-\theta(z)s}\} a(s,z)\,\mathrm{d}s\,\mathrm{d}A_{0,c}(z) \right].$$

Proof. This can be proved following the lines of lemma A.3 in Hjort (1990b), starting from the following argument: If $E\exp(-\theta Y) = \exp\{-h(\theta)\}$, then $E(1 - rY)\exp(-\theta Y) = \{1 - rh'(\theta)\}\exp\{-h(\theta)\}$.

The survival data set of (t_i, δ_i, x_i) can be represented by the triples (N_i, Y_i, x_i). It will be convenient to refer to the full data set as $\mathbf{D}_n$, and, as we need to derive results via induction, to (N_i, Y_i, x_i) data about the k first individuals as $\mathbf{D}_k$. We first address the simplest case of only one observation, where lemma 2 gives the prior-to-posterior results in terms of the Lévy measure $\nu(\mathrm{d}s, \mathrm{d}t)$.

Lemma 2
Let A be a Beta process (c, A_0), with A_0 having jumps in $\mathcal{M} = \{t_1, \ldots, t_m\}$. For given γ, the conditional distribution of A given $\mathbf{D}_1$ is a Lévy process with Lévy measure $\nu(\mathrm{d}s, \mathrm{d}t \mid \gamma, \mathbf{D}_1)$ equal to

$$\{1 - r(x_1^{\mathrm{t}}\gamma)s\}^{Y_1(t)} a(s,t)\,\mathrm{d}s\,\mathrm{d}A_{0,c}(t)$$
$$+ \sum_{j=1}^m k_j^{-1} \{1 - r(x_1^{\mathrm{t}}\gamma)s\}^{1 - I\{\Delta N_1(t_j) = 1\}} \{r(x_1^{\mathrm{t}}\gamma)s\}^{I\{\Delta N_1(t_j) = 1\}} f_{t_j}(s)\,\mathrm{d}s\,\delta_{t_j}(\mathrm{d}t)$$
$$+ k(t)^{-1} r(x_1^{\mathrm{t}}\gamma) s a(s,t)\,\mathrm{d}s\,\mathrm{d}N_1^*(t),$$

where k_j and $k(t)$ are normalizing constants, $f_{t_j}(s)$ is the Beta density given by (11) and $N_1^*(t) = N_1(t) - \Delta N_1(t) I\{t \in \mathcal{M}\}$.

Proof. The proof follows the arguments used by Hjort (1990b, theorem 4.1) in concert with lemma 1. It requires working separately with three different cases, namely the case of a censored observation and the cases of a complete observation at a time point either outside or inside $\mathcal{M}$.

Next, we derive the posterior of A in terms of the Lévy measure of the continuous part and the density of the jumps. For this we need some further notation. We define $R_{n,t}(s,\gamma)$ as $\prod_{j \in R_n(t)}\{1 - r(x_j^{\mathrm{t}}\gamma)s\}$, where $R_n(t)$ is the risk set $\{i : Y_i(t) = 1\}$. Similarly, we define the index set $R_n^+(t) = \{i : Y_i(t) - \Delta N_i(t) = 1\}$ of individuals still at risk at t, apart from the individual who dies at t, and the corresponding quantity $R_{n,t}^+(s,\gamma)$.

Theorem 1
Let A be a Beta process (c, A_0) with a continuous A_0 (i.e. $\mathcal{M} = \emptyset$). Then, conditional on γ and $\mathbf{D}_n$, the posterior distribution of A is a Lévy process with Lévy measure

$$\nu(\mathrm{d}s, \mathrm{d}t \mid \gamma, \mathbf{D}_n) = R_{n,t}(s,\gamma) a(s,t)\,\mathrm{d}s\,\mathrm{d}A_0(t) + \sum_{i=1}^{q_n} h_i(s,\gamma)\,\mathrm{d}s\,\delta_{t_i}(\mathrm{d}t), \tag{19}$$

where

$$h_i(s,\gamma)=\left[\prod_{j=1}^{n}\{r(x_j^t\gamma)s\}^{\Delta N_j(t_i)}\right]R^+_{n,t_i}(s,\gamma)a(s,t_i)/k_i(\gamma) \quad \textit{for } s\in(0,1) \tag{20}$$

is the jump density at t_i, with $k_i(\gamma)$ the normalizing constant.

Proof. For $n=1$, lemma 2 with $\mathcal{M}=\emptyset$ implies (19). For $n\geq 2$ the proof is completed by applying lemma 2 repeatedly.

If there are no ties among complete observations, equation (20) simplifies to

$$h_i(s,\gamma)=r(x_i^{\mathrm{t}}\gamma)R^+_{n,t_i}(s,\gamma)c(t_i)(1-s)^{c(t_i)-1}/k_i(\gamma) \quad \text{on } (0,1). \tag{21}$$

According to formula (15), the posterior expectation $\hat{A}(t,\gamma)=E\{A(t)\,|\,\gamma,\mathbf{D}_n\}$ takes the form

$$\int_0^t\int_0^1 sv(\mathrm{d}s,\,\mathrm{d}z\,|\,\gamma,\mathbf{N}_n)=\mathrm{one}_n+\mathrm{two}_n, \tag{22}$$

say, where the two terms on the right-hand side can be represented as $\int_0^t\int_0^1 sR_{n,z}(s,\gamma)\times a(s,z)\,\mathrm{d}s\,\mathrm{d}A_0(z)$ and $\sum_{i=1}^{q_n}I\{t_i\leq t\}\int_0^1 sh_i(s,\gamma)\,\mathrm{d}s$, respectively. For computing these, let us define the integral

$$J^I(a,b)=\int_0^1 s^{a-1}(1-s)^{b-1}\prod_{j\in I}\{1-r(x_j^{\mathrm{t}}\gamma)s\}\,\mathrm{d}s, \tag{23}$$

where I is a subset of the first n integers. We can then check that

$$\begin{aligned}\mathrm{one}_n&=\int_0^t J^{R_n(z)}(1,c(z))c(z)\,\mathrm{d}A_0(z),\\ \mathrm{two}_n&=\sum_{i=1}^{q_n}I\{t_i\leq t\}\frac{J^{R_n^+(t_i)}(\mathrm{d}N_\cdot(t_i)+1,c(t_i))}{J^{R_n^+(t_i)}(\mathrm{d}N_\cdot(t_i),c(t_i))}.\end{aligned} \tag{24}$$

5.2. Marginal posterior distribution of γ

In order to derive the marginal posterior distribution of γ we proceed as follows: first, we derive the marginal compensator of the counting process N_{k+1}, given γ and $\mathbf{D}_k$; secondly, we derive the likelihood of γ using Jacod's formula for the likelihood ratio. Unless otherwise specified, r_i below will be short-hand notation for $r(x_i^{\mathrm{t}}\gamma)$, for $i=1,\ldots,n$.

Lemma 3

For $k\geq 1$, conditionally on γ and $\mathbf{D}_k$, N_{k+1} is a multiplicative counting process with compensator $\int_0^t Y_{k+1}(z)\,\mathrm{d}E\{A(z,x_{k+1})\,|\,\gamma,\mathbf{D}_k\}$, where

$$E\{A(t,x_{k+1})\,|\,\gamma,\mathbf{D}_k\}=\int_0^{\mathrm{t}}\int_0^1 r_{k+1}s\,v(\mathrm{d}s,\,\mathrm{d}z\,|\,\gamma,\mathbf{D}_k). \tag{25}$$

Proof. Theorem 1 implies that, conditional on γ and $\mathbf{D}_k$, the posterior distribution of A is a Lévy process with Lévy measure $v(\mathrm{d}t,\,\mathrm{d}s\,|\,\gamma,\mathbf{D}_k)$. We find that $\Pr\{T_{k+1}^0>t\,|\,\gamma,\mathbf{D}_k\}$ may be written

$$E\left[\prod_{[0,t]}\{1-\mathrm{d}A(s,x_{k+1})\}\Big|\gamma,\mathbf{D}_k\right]=\prod_{[0,t]}\{1-\mathrm{dE}[A(s,x_{k+1})\,|\,\gamma,\mathbf{D}_k]\}.$$

Hence, conditional on γ and $\mathbf{D}_k$, the cumulative hazard of T_{k+1}^0 is $E\{A(s, x_{k+1}) \mid \gamma, \mathbf{D}_k\}$. Similar to (15), $E\{A(t) \mid \gamma, \mathbf{D}_k\}$ can be expressed as $\int_{[0,t]\times[0,1]} s\nu(\mathrm{d}s, \mathrm{d}z \mid \gamma, \mathbf{D}_k)$, and $E\{A(s, x_{k+1}) \mid \gamma, \mathbf{D}_k\}$ may be recovered by simply multiplying for r_{k+1}.

Lemma 4

Define the likelihood ratio

$$L_{k+1}(N_{k+1} \mid \gamma, \mathbf{D}_k) = \frac{\mathrm{d}P_{k+1}(\cdot \mid \gamma, \mathbf{D}_k)}{\mathrm{d}P_{k+1}(\cdot \mid 0, \mathbf{D}_k)},$$

where $P_{k+1}(\cdot \mid \gamma, \mathbf{D}_k)$ is the probability measure of N_{k+1} conditional on γ and $\mathbf{D}_k$. This likelihood ratio is then proportional to

$$\begin{aligned}
&\prod_{i=1}^{q_k} \left[k_i^{-1} \int_0^1 (1 - r_{k+1}s)^{1-\Delta N_{k+1}(t_i)} (r_{k+1}s)^{\Delta N_{k+1}(t_i)} \right. \\
&\quad \left. \times \prod_{j=1}^{k} (r_j s)^{\Delta N_j(t_i)} (1 - r_j s)^{Y_j(t_i) - \Delta N_j(t_i)} a(s, t_i)\, \mathrm{d}s \right]^{Y_{k+1}(t_i)} \\
&\quad \times \exp\left\{ - \int_0^{T_{k+1}} \int_0^1 r_{k+1} s \prod_{j=1}^{k} (1 - r_j s)^{Y_j(t)} a(s, t)\, \mathrm{d}s\, \mathrm{d}A_0(t) \right\} \\
&\quad \times \left[\int_0^1 r_{k+1} s \prod_{j=1}^{k} (1 - r_j s)^{Y_j(T_{k+1})} a(s, T_{k+1})\, \mathrm{d}s \right]^{I\{\delta_{k+1}=1,\, T_{k+1} \neq t_i,\, i=1,\ldots,q_k\}},
\end{aligned}$$

where q_k and $(t_1, \ldots, t_{q_k})$ are the parallels of q_n and $(t_1, \ldots, t_{q_n})$ with respect to the first k observations, and

$$k_i = \int_0^1 \prod_{j=1}^{k} (r_j s)^{\Delta N_j(t_i)} (1 - r_j s)^{Y_j(t_i) - \Delta N_j(t_i)} a(s, t_i)\, \mathrm{d}s. \tag{26}$$

Proof. The proof partly follows the lines of lemma 5.3 of Kim & Lee (2003a). The idea is to distinguish in $E[A(s, x_{k+1}) \mid \gamma, \mathbf{D}_k]$ between the continuous part and the discrete part and then use Jacod's formula (see ABGK, section II.7). For $B(t) = E\{A(s, x_{k+1}) \mid \gamma, \mathbf{D}_k\}$, define $B_d(t) = \sum_{i=1}^{q_k} \Delta B(t_i) I\{t_i \le t\}$ and $B_c(t) = B(t) - B_d(t)$. Then we have

$$\begin{aligned}
L_{k+1}(N_{k+1} \mid \gamma, \mathbf{D}_k) &\propto \prod_t \mathrm{d}B(t)^{\Delta N_{k+1}(t)} \{1 - Y_{k+1}(t)\, \mathrm{d}B(t)\}^{1 - \Delta N_{k+1}(t)} \\
&= b_c(T_{k+1})^{I\{\delta_{k+1}=1,\, T_{k+1} \neq t_i,\, i=1,\ldots,q_k\}} \exp\left\{ - \int_0^\infty Y_{k+1}(t)\, \mathrm{d}B_c(t) \right\} \\
&\quad \times \prod_{i=1}^{q_k} \Delta B_d(t_i)^{\Delta N_{k+1}(t_i)} \{1 - Y_{k+1}(t_i) \Delta B_d(t_i)\}^{1 - \Delta N_{k+1}(t_i)},
\end{aligned}$$

where $b_c(t)$ is the first derivative of $B_c(t)$. From the definition of $B(t)$ in (25), we may write $B_d(t)$ as

$$\sum_{i=1}^{q_k} k_i^{-1} \int_0^1 r_{k+1} s \prod_{j=1}^{k} (r_j s)^{\Delta N_j(t_i)} (1 - r_j s)^{Y_j(t_i) - \Delta N_j(t_i)} a(s, t_i)\, \mathrm{d}s\, I\{t_i \le t\}$$

whereas

$$B_c(t)=\int_0^t\int_0^1 r_{k+1}s\prod_{j=1}^{k}(1-r_js)^{Y_j(z)}a(s,z)\,\mathrm{d}s\,\mathrm{d}A_0(z).$$

The proof is completed by substituting these expressions into the L_{k+1} formula above.

Theorem 2
Let $A\sim$ Beta (c,A_0) with continuous A_0 and let γ, independently of A, have density $\pi(\gamma)$. Then the marginal posterior distribution of γ given $\mathbf{D}_n$ is given by

$$\pi(\gamma\,|\,\mathbf{D}_n)\propto\pi(\gamma)\exp\{-\rho_n(\gamma)\}\prod_{i=1}^{q_n}k_i(\gamma), \tag{27}$$

where

$$\rho_n(\gamma)=\sum_{i=1}^{n}\int_0^{T_i}\int_0^1 r(x_i^{\mathrm{t}}\gamma)s\prod_{j=1}^{i-1}\{1-r(x_j^t\gamma)s\}^{Y_j(t)}a(s,t)\,\mathrm{d}s\,\mathrm{d}A_0(t), \tag{28}$$

and $k_i(\gamma)$ is the normalizing constant of (20).

Proof. When $n=1$, (27) is a consequence of lemma 4. Assume that (27) is true for $n=k$. As

$$\pi(\gamma\,|\,\mathbf{D}_{k+1})\propto L_{k+1}(N_{k+1}\,|\,\gamma,\mathbf{D}_k)\pi(\gamma\,|\,\mathbf{D}_k),$$

we have verified (27) for $n=k+1$, by rearranging the equation $L_{k+1}(N_{k+1}\,|\,\gamma,\mathbf{D}_k)$ in lemma 4 and $\pi(\gamma\,|\,\mathbf{D}_k)$ in (27). Note that the normalizing constants k_i in (26) simplify.

We may arrange $t_1\le\cdots\le t_n$, without loss of generality. Then, by rearranging the integrals in (28) and using $a(s,t)$ of (13), we find

$$\rho_n(\gamma)=\sum_{i=1}^{n}\int_0^{T_i}\int_0^1 r(x_i^{\mathrm{t}}\gamma)\prod_{j=i+1}^{n}\{1-r(x_j^{\mathrm{t}}\gamma)s\}c(t)(1-s)^{c(t)-1}\,\mathrm{d}s\,\mathrm{d}A_0(t).$$

Remark 1. Theorems 1 and 2 together provide a full description of the semiparametric posterior distribution of (A,γ), valid for any Beta process with positive $c(\cdot)$ function. When $c(\cdot)$ is large compared with sample size n, A is close to A_0 with high probability also in the posterior distribution, and Bayes inference approximates the parametric methods associated with a fixed A_0 and unknown γ. Using $c(\cdot)$ small, compared with n, on the other hand, has notable consequences for the posterior. We see from (13) and results above that the $\rho_n(\gamma)$ becomes close to zero, meaning that the posterior for γ is close to being proportional to $\prod_{i=1}^{q_n}k_i(\gamma)$, which can be approximated further via results from our appendix. Also, the distribution of A given γ becomes mostly concentrated in only the jumps at the observed life times, i.e. the continuous part disappears. In particular, when $c(\cdot)\to 0$, the one$_n$ of (22) and (24) disappears.

6. Simulating from the posterior distribution

Here, we develop a Markov Chain Monte Carlo algorithm for the Bayesian analysis of the proportional regression model (9). The parameters determining the model are the baseline

cumulative hazard function, A, and the regression coefficient vector, γ. We wish to make inference for various quantities that depend on (A, γ), like $A(t;x) = A(t)r(x^{\mathrm{t}}\gamma)$, the cumulative hazard of an individual with a certain covariate value, or the survival probability $\Pr\{T > t \mid x\} = \pi_{[0,t]}\{1 - \mathrm{d}A(s)r(x^{\mathrm{t}}\gamma)\}$. Another quantity of immediate interest is the *median residual life length* for a person with covariate x who has survived up to time t_0 (see e.g. Gould, 1995):

$$\mathrm{med}(t_0;x) = B^{-1}\left(\frac{\log 2}{r(x^{\mathrm{t}}\gamma)}\right) - t_0, \qquad \text{where } B(t) = A(t) - A(t_0) \text{ for } t \geq t_0. \tag{29}$$

One strategy for estimating $A(t;x)$, from a similar recipe in Hjort (1990b, section 6), is to first generate samples from the posterior of γ, and then to compute the semiparametric estimator by posterior averaging, using formulae for one_n and two_n in connection with (22) and (24):

$$\hat{A}(t;x) = E\left[E\{A(t;x) \mid \gamma, \mathrm{data}\} \mid \mathrm{data}\right] = \int r(x^{\mathrm{t}}\gamma)\hat{A}(t,\gamma)\pi(\gamma \mid \mathrm{data})\,\mathrm{d}\gamma.$$

The problem is that $\pi(\gamma \mid \mathrm{data})$ is proportional to a term containing integrals, of the form (23). Exact formulae for these may be found, in terms of finitely many Beta functions, but the number of required terms quickly becomes prohibitively large; this also spells trouble for Metropolis–Hasting samplers with ratios of such integrals as acceptance ratios. Instead we aim at simulating from the joint posterior $[A, \gamma \mid \mathrm{data}]$, making it possible to draw inference on $A(t;x)$, $\mathrm{med}(t_0;x)$ and other quantities by simply plugging in the corresponding (A, γ) samples. We shall develop a Gibbs sampler that generates iteratively from $[A \mid \gamma, \mathrm{data}]$ and $[\gamma \mid A, \mathrm{data}]$. Our scheme resembles recipes worked with in Laud *et al.* (1998) and Lee & Kim (2004) for handling the numerical side of Hjort's (1990b) semiparametric Bayes methods for the classical Cox model.

For given γ, the posterior of A can be decomposed into its continuous part A_{c} and discrete part A_{d}; $A = A_{\mathrm{c}} + A_{\mathrm{d}}$. Here A_c is a Lévy process with Poisson mean measure $\nu_n(\mathrm{d}s, \mathrm{d}t) = R_{n,t}(s,\gamma)a(s,t)\,\mathrm{d}s\,\mathrm{d}A_0(t)$, while A_{d} has only fixed jumps at non-censored data points $t_1, \ldots, t_{q_n}$. For simulating sample paths of A_{c} there are in essence two approaches associated with the machinery of Lévy processes. The first is to sample the increments of A_c according to a partition of the time axis, usually with respect to distinct data points. For consecutive time points t_{i-1} and t_i, $A_{\mathrm{c}}(t_i) - A_{\mathrm{c}}(t_{i-1})$ has an infinitely divisible distribution, and sampling algorithms are available in the literature (cf. Ferguson & Klass, 1972; Damien *et al.*, 1995; Wolpert & Ickstadt, 1998; see Walker & Damien, 2000, for a recent discussion).

However, we prefer the second approach which is to approximate A_{c} by a cumulative compound point process. The following *Poisson weighted algorithm* extends ideas present in Damien *et al.* (1996) and utilized more fully in Hjort & Kim (2007). The algorithm can be described as follows. For a Lévy process A with Poisson mean measure ν on $[0,1] \times [0,\tau]$, where $\nu(\mathrm{d}s, \mathrm{d}t) = s^{-1}f_t(s)\,\mathrm{d}s\,\mathrm{d}A_0(t)$, let, for each $t \in [0,\tau]$, $g_t(s)$ be a probability density on $[0,1]$, and suppose it allows easy simulation. (i) For a large m, generate $z_1, \ldots, z_m$ from $\mathrm{d}A_0(t)/A_0(\tau)$; (ii) for $i = 1, \ldots, m$, generate $s_i \sim g_{z_i}(s)$; (iii) for $i = 1, \ldots, m$, set $\lambda_i = A_0(\tau)f_{z_i}(s_i)/\{ms_i g_{z_i}(s_i)\}$, then generate $w_i \sim \mathrm{Pois}(\lambda_i)$; and (iv) set $A_m(t) = \sum_{i=1}^{m} s_i w_i I\{z_i \leq t\}$.

Theorem 3

The process A_m generated by the Poisson weighted algorithm converges in distribution to a Lévy process with mean measure $\nu(\mathrm{d}s, \mathrm{d}t) = s^{-1}f_t(s)\,\mathrm{d}s\,\mathrm{d}A_0(t)$, on the space $D[0,\tau]$ of cadlag function on $[0,\tau]$ endowed with the Skorohod topology.

Proof. Let $z \sim \mathrm{d}A_0(t)/A_0(\tau)$, then $[s \mid z] \sim g_z$, and let $[w \mid z, s]$ be Poisson with parameter $A_0(\tau)f_z(s)/\{msg_z(s)\}$. For a piecewise constant function $\theta(t)$ on $[0, \tau]$, consider the Laplace transform

$$E \exp\left\{-\int_0^\tau \theta(t)\,\mathrm{d}A_m(t)\right\} = E \exp\left\{-\sum_{i=1}^m \theta(z_i)s_i w_i\right\} = \left[E \exp\{-\theta(z)sw\}\right]^m.$$

By conditional expectation, $E \exp\{-\theta(z)sw\}$ can be written

$$EE[\exp\{-\theta(z)sw\} \mid (s,z)] = E \exp\left(-\frac{A_0(\tau)f_z(s)}{msg_z(s)}\left[1 - \exp\{-s\theta(z)\}\right]\right),$$

where the last expectation is with respect to the distribution of (s, z).

Next let us write $B = [1 - \exp\{-s\theta(z)\}]A_0(\tau)f_z(s)/\{sg_z(s)\}$, which is seen to be bounded. Then we may state that, as $m \to \infty$,

$$\left\{E \exp\left(-\frac{B}{m}\right)\right\}^m = \{E(1 - m^{-1}B + (1/2)m^{-2}B^2 - \cdots)\}^m \to \exp(-EB).$$

For the case at hand,

$$\begin{aligned} EB &= \int_0^\tau \int_0^1 (1 - \mathrm{e}^{-s\theta(t)}) \frac{A_0(\tau)f_t(s)}{sg_t(s)}\, g_t(s)\,\mathrm{d}s\, \frac{\mathrm{d}A_0(t)}{A_0(\tau)} \\ &= \int_0^\tau \int_0^1 (1 - \mathrm{e}^{-s\theta(t)}) s^{-1} f_t(s)\,\mathrm{d}s\,\mathrm{d}A_0(t). \end{aligned}$$

Thus, the log-Laplace functional of $A_m(t)$ converges to $-\int_0^\tau \int_0^1 (1 - \mathrm{e}^{-s\theta(t)}) s^{-1} f_t(s)\,\mathrm{d}s\,\mathrm{d}A_0(t)$, which corresponds to a Lévy process with Poisson mean measure $\nu(\mathrm{d}s, \mathrm{d}t) = s^{-1}f_t(s)\,\mathrm{d}s\,\mathrm{d}A_0(t)$, completing the proof.

As for the choice of $g_t(s)$, two different densities have been investigated. The first idea is to set $g_t(s) = \mathrm{beta}(s; 1, c(t))$. Simulation experiences have shown that the algorithm then fails to catch a sufficient amount of the smaller jumps, in that many of the Poisson variates become zero. A better numerical performance is typically achieved with $g_t(s) = \mathrm{beta}(s; 1, c(t) + \sum_{j \in R_n(t)} r(x_j^{\mathrm{t}}\gamma))$, with corresponding formula for λ_i,

$$\lambda_i = \frac{A_0(\tau)}{ms_i} \frac{c(z_i)}{c(z_i) + \sum_{j \in R_n(z_i)} r(x_j^{\mathrm{t}}\gamma)} \frac{R_{n,z_i}(s_i, \gamma)}{(1 - s_i)^{\sum_{j \in R_n(z_i)} r(x_j^{\mathrm{t}}\gamma)}}.$$

For A_d, we use *Gibbs step with auxiliary variables* for generating from a density proportional to $h_i(s, \gamma)$ in (20), for each fixed jump at $t_1, \ldots, t_{q_n}$. In particular, we use a sampling scheme of Besag & Green (1993). Consider in general a density $\pi(s)$ proportional to $\prod_{i=0}^k f_i(s)$. We describe a Markov chain, s_t, that passes from a current iteration $s_t = s$ to the next position $s_{t+1} = s^*$. For such a given jump size s, first generate k independent uniform variables $u_i \sim \mathrm{unif}[0, f_i(s)]$ for $i = 1, \ldots, k$. Then, given $(u_1, \ldots, u_k)$, generate s^* from a density proportional to $f_0(s)I\{s : f_i(s) \geq u_i \text{ for } i = 1, \ldots, k\}$. The marginal density of such an s^* is then proportional to $f_0(s), \ldots f_k(s)$.

The density of the jump at t_i is given by $[s_i] \propto \prod_{j \in R_n^+(t_i)} \{1 - r(x_j^{\mathrm{t}}\gamma)s_i\}(1 - s_i)^{c(t_i)-1}$. We simulate from this as follows: (i) for $j \in R_n^+(t_i)$, generate $u_{i,j} \sim$ unif $[0, 1 - r(x_j^{\mathrm{t}}\gamma)s_i]$; (ii) generate s_i^* from the Beta$(1, c(t_i))$ density, but truncated to $s \leq \min_{j \in R_n^+(t_i)}\{(1 - u_{i,j})/r(x_j^{\mathrm{t}}\gamma), 1\}$, which is accomplished using the raw inverse cumulative distribution function method. Thus we end up working with the joint density

$$[s_i, \mathbf{u}_i] \propto \prod_{j \in R_n^+(t_i)} I\left\{u_{i,j} \leq 1 - r(x_j^{\mathrm{t}}\gamma)s_i\right\}(1 - s_i)^{c(t_i)-1}.$$

For a given path of A, γ can be sampled from the corresponding full conditional as follows. Let $s_1^*, \ldots, s_{q_n}^*$ be the jump sizes of $A_{\rm d}$ at $t_1, \ldots, t_{q_n}$ and let $(z_1, s_1), \ldots, (z_m, s_m)$ be the jump times and the jump sizes of $A_{\rm c}$. For each $t \in [0, \tau]$,

$$A(t) = \sum_{i=1}^{m} s_i I\{z_i \le t\} + \sum_{i=1}^{q_n} s_i^* I\{t_i \le t\}. \tag{30}$$

Adapting the likelihood (18) to the notation above, the posterior of γ given A and data are given by

$$[\gamma \mid A, \text{data}] \propto \pi(\gamma) \prod_{i=1}^{q_n} \{r(x_i^{\rm t}\gamma) s_i^* R^+_{n,\,t_i}(s_i^*, \gamma)\} \times \prod_{i=1}^{m} R_{n,\,z_i}(s_i, \gamma). \tag{31}$$

We incorporate the conditionals of the variables $\mathbf{U} = \{\mathbf{u}_i,\ i = 1, \ldots, q_n\}$ to get

$$[\gamma \mid A, \mathbf{U},\ \text{data}\ ] \propto \pi(\gamma) \prod_{i=1}^{q_n} r(x_i^{\rm t}\gamma) \prod_{j \in R_n^+(t_i)} I\left\{ r(x_j^{\rm t}\gamma) \le \frac{1 - u_{i,j}}{s_i^*} \right\} \times \prod_{i=1}^{m} R_{n,\,z_i}(s_i, \gamma).$$

The γ is hence restricted to the subset of $\mathcal{R}^p$ satisfying the product of indicators and we may implement a Metropolis–Hastings step with no particular difficulties. We use a random walk kernel as proposal distribution.

The full Gibbs sampler algorithm for the augmented space $(A, \mathbf{U}, \gamma)$ can be summarized as follows, giving realizations of $A = A_{\rm c} + A_{\rm d}$ as in (30). *Part 1:* for $[A_c \mid \gamma, \text{data}]$, use the Poisson weighted algorithm: (i) for large m, generate $z_1, \ldots, z_m$ from ${\rm d}A_0(t)/A_0(\tau)$, and sort them in increasing order; (ii) for $i = 1, \ldots, m$, generate $s_i' \sim \text{Beta}\{1, c(z_i) + \sum_{j \in R_n(z_i)} r(x_j^{\rm t}\gamma)\}$; (iii) for $i = 1, \ldots, m$ set

$$\lambda_i = \frac{A_0(\tau)}{m s_i'} \frac{c(z_i)}{c(z_i) + \sum_{j \in R_n(z_i)} r(x_j^{\rm t}\gamma)} \frac{R_{n,\,z_i}(s_i', \gamma)}{(1 - s_i')^{\sum_{j \in R_n(z_i)} r(x_j^{\rm t}\gamma)}}$$

and then generate $w_i \sim \text{Pois}(\lambda_i)$; (iv) for $i = 1, \ldots, m$ set $s_i = s_i' w_i$; and (v) finally form $A_c(t) = \sum_{i=1}^{m} s_i I\{z_i \le t\}$. *Part 2:* for $[A_d, \mathbf{U} \mid \gamma, \text{data}]$, use a Gibbs step: for $i = 1, \ldots, q_n$, (i) for $j \in R_n^+(t_i)$, generate $u_{i,j} \sim \text{unif}[0, 1 - r(x_j^{\rm t}\gamma) s_{i(-1)}^*]$, where $s_{i(-1)}^*$ is the value at the previous iteration; (ii) generate s_i^* from the Beta density $(1, c(t_i))$, but truncated to $s \le \min_{j \in R_n^+(t_i)}\{(1 - u_{i,j})/r(x_j^{\rm t}\gamma), 1\}$, and then form the second sum of (30). *Part 3:* for $[\gamma \mid A, \mathbf{U}, \text{data}]$, use a random walk Metropolis–Hastings step: let γ' be a candidate value for γ generated by a random walk kernel $q(\gamma, \gamma')$. Then, the acceptance ratio is

$$\frac{\pi(\gamma')}{\pi(\gamma)} \frac{\prod_{i=1}^{q_n} r(x_i^{\rm t}\gamma') \times \prod_{i=1}^{m} R_{n,\,z_i}(s_i, \gamma')}{\prod_{i=1}^{q_n} r(x_i^{\rm t}\gamma) \times \prod_{i=1}^{m} R_{n,\,z_i}(s_i, \gamma)} I\{\gamma' \in B\},$$

where $\pi(\gamma)$ is the Jeffreys prior identified in section 4 and B is the set of γ' for which $r(x_j^{\rm t}\gamma') \le (1 - u_{i,j})/s_i^*; i = 1, \ldots, q_n, j \in R_n^+(t_i)$; note that γ at the previous iteration belongs to B, by construction.

7. Illustration

As illustration we consider a data set from a Danish medical study for the period 1962–77. It concerns 205 patients with malignant melanoma who had a radical operation for removing the tumour and were followed till death. Time is measured since operation, in days, which we convert to a time scale of years in our illustrations below, together with a categorical variable, which records the type of event which has occurred. There are 57 deaths as a result of melanoma, 134 censored observations, and a further 14 deaths due to other causes, which we lump together with the censored ones. Among various explanatory variables we

decided to include the thickness of the tumour, expressed in mm after subtracting the sample mean (2.92 mm). We use these data for two different types of illustration. First, we give a partial likelihood-based analysis of the model (2), where we evaluate its behaviour in comparison with the Cox model (1). Afterwards we use the data set to illustrate Bayesian inference, with focus on the median residual life-length parameter for a person with given covariate x upon survival at time t_0, i.e. $\text{med}(t_0, x)$ of (29).

7.1. Comparison between the two models

Partial likelihood inference on the regression coefficient gives a large-sample 95% confidence interval of $[0.607, 1.355]$, based on the maximum partial likelihood estimate $\hat{\gamma} = 0.981$ with observed information matrix $-n^{-1}\ell_n^{(2)}(\hat{\gamma}, \tau) = 0.134$, the usual information calculus operation on the log-partial likelihood ℓ_n of (4). We also compute corresponding estimates for the Cox model. To check adequacy of both the models (1) and (2), we may compute the estimated cumulative hazard rates

$$\hat{Z}_i = \hat{A}_{\text{Cox}}(t_i, \hat{\beta}) \exp(x_i^{\text{t}}\hat{\beta}) \quad \text{and} \quad Z_i^* = \hat{A}(t_i, \hat{\gamma}) \frac{\exp(x_i^{\text{t}}\hat{\gamma})}{1 + \exp(x_i^{\text{t}}\hat{\gamma})}, \tag{32}$$

featuring the Breslow–Aalen estimators in the Cox model and the logistic model respectively. Under Cox model conditions the $\hat{Z}_i$s mimic a sample from the standard exponential, and similarly under the logistic model it is the Z_i^*s that best should mimic such a sample. Figure 1 (left panel) displays Nelson–Aalen plots for the quantities of (32). The plots suggest a preference for model (2). It may be interpreted as a lack of fit of the Cox model in capturing long survival and high thickness, as thickness has a positive effect on the hazard; as pointed out earlier, the Cox model sometimes runs the risk of overestimating the relative risk for individuals with large covariate values. We also fitted the extended model

$$\alpha(s, x) = \alpha(s) \frac{\exp(x^{\text{t}}\gamma)}{\{1 + \exp(x^{\text{t}}\gamma)\}^{\kappa}} \tag{33}$$

via partial likelihood, finding $\hat{\kappa} = 1.008$, see the profile log-partial likelihood of Fig. 1 (right panel). A 95% confidence interval, based on profile deviance, is found to be $[0.873, 1.118]$, clearly ruling out the case $\kappa = 0$, i.e. the Cox model.

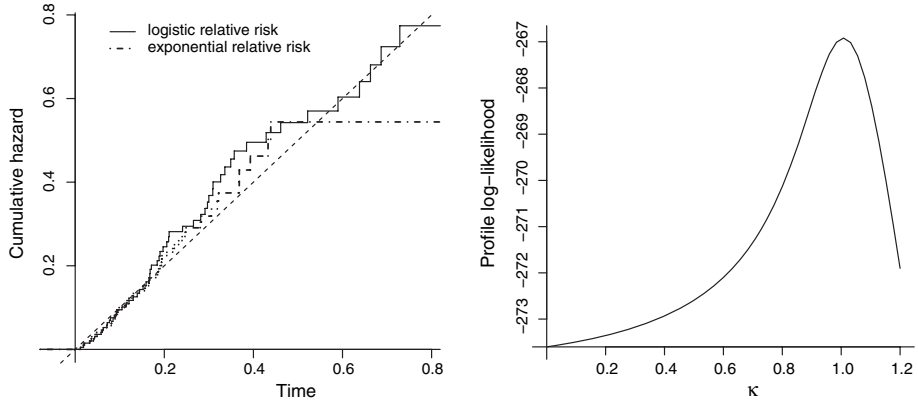

Fig. 1. For the Danish melanoma data, with $n = 205$ patients, and tumour thickness as covariate, the left panel displays Nelson–Aalen plots of $\hat{Z}_i$ and Z_i^* of (32) for models (2) and (1), and the right panel shows the profile log-partial likelihood for the (33) model, with estimate $\hat{\kappa} = 1.008$.

ABGK (Section VII.3) used the melanoma data to illustrate goodness-of-fit methods for the Cox regression model. In particular, they found that thickness does not satisfy the log-linearity hypothesis, showing that the influence on the log-hazard seems to be a concave function, arguing therefore that thickness might better enter the Cox model after a logarithmic transformation. Our analysis indicates that a logistic relative risk function represents a viable alternative to the log-transformation for describing the effect of thickness on the hazard.

7.2. Bayesian inference for median remaining lifetime

For a Bayes illustration with the melanoma data we use a Beta process with prior parameters $A_0(t)=a_0t$ and $c(t)=k\exp(-a_0t)$. It corresponds to a baseline hazard function of the exponential type, where the k governs the strength of prior beliefs. Such a k is close to having a prior sample size interpretation. In fact, with i.i.d. survival times, $c(t)=k\exp(-a_0t)$ may be interpreted as the number at risk at t in an imagined prior sample of uncensored survival times, with k the sample size. We may compare k with n, the number of observations in the data, for assessing the strength of prior beliefs. We ran the algorithm for different values of a_0 and k so as to investigate the sensitivity of Bayes analysis to the prior. Here, we show the result for $a_0=0.0475$ and $k=10$. For the Metropolis–Hastings step we set $q(\gamma,\gamma')=\mathrm{N}(\gamma';\gamma,\sigma)$ with $\sigma=1$. For the compound point process that approximates the continuous part of the posterior of A we set $m=500$, a sufficiently large value with respect to the average number of non-zero Poisson variates we observed in pilot runs of the algorithm with varying m. The Gibbs sampler was run for 100,000 iterations, and we discarded the first 10,000 as burn-in. The analysis is therefore carried out with a total of 90,000 samples.

Sampling from the posterior distribution of γ gives a credibility interval of [0.787, 1.639], shifted to the right with respect to traditional likelihood inference. In the left panel of Fig. 2 we show the histogram of the posterior sample together with the approximating normal distribution implied by frequentist large-sample theory, as in section 3. The right panel shows the pointwise 90% credible band, together with the posterior median obtained from the simulated paths. For comparison with partial likelihood estimation, we add the Breslow–Aalen estimator. We may note that the estimate of the baseline based on data is higher than the

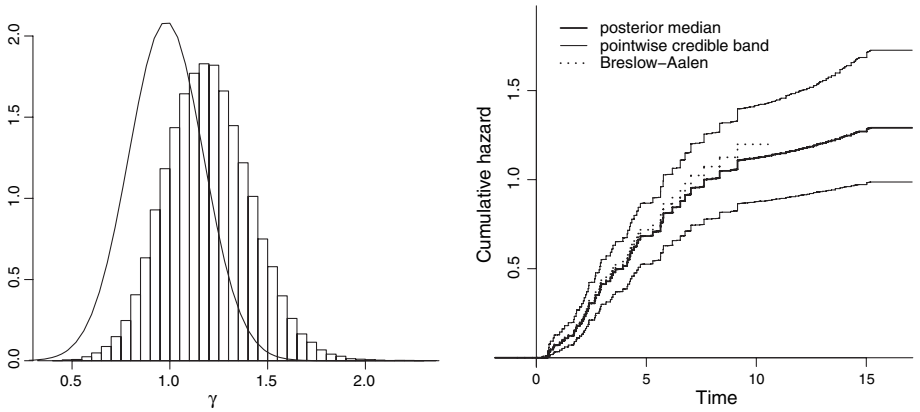

Fig. 2. For the Danish melanoma data, with tumour thickness as covariate: the left panel shows the histogram of posterior samples of γ against the large-sample approximation $N(\hat{\gamma},n^{-1}\hat{\Sigma}^{-1})$. The right panel shows a pointwise 90% credibility band for the baseline cumulative hazard A along with the Breslow–Aalen estimator, with years as time scale. The simulation sample size from the posterior is 90,000.

Table 1. *Posterior distribution summary for the median remaining life length, for a patient with thickness equal to 3.92 mm, and who has survived, respectively, 0, 1 and 2 years since operation*

Lifetime (years)	Residual posterior mean	Residual 90% credible intervals	Absolute 90% credible intervals
$t_0 = 0$	6.76	5.30–11.76	5.30–11.76
$t_0 = 1$	6.62	4.76–12.30	5.76–13.30
$t_0 = 2$	7.15	4.54–12.38	6.54–14.38

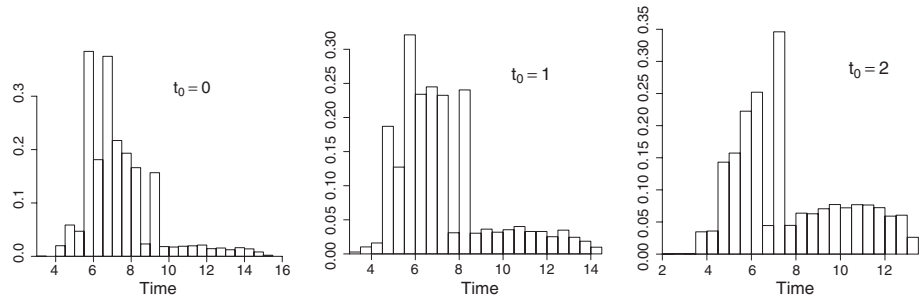

Fig. 3. How long one can expect to survive further, when one already has survived a certain period after operation? For a patient with $x = 1$, corresponding to tumour thickness 3.92 mm, the figure displays the posterior distribution of the median remaining lifetime just after operation (left panel), after 1 year of survival (central panel) and after 2 years of survival (right panel). The time unit is years.

posterior pointwise median. Actually a_0 is about the same size as the maximum partial likelihood estimate of the hazard rate under the hypothesis of i.i.d. exponential survival times, roughly speaking the hazard level of an individual in the mean, hence about one-half of the baseline hazard displayed by the data.

Concerning the median remaining lifetime quantity $\text{med}(t_0; x)$ of (29), for a patient with covariate x who has already survived a time period t_0, Table 1 reports on the posterior distribution of this quantity. The covariate is $x = 1$ (corresponding to tumour thickness 3.92 mm), and our imagined patient has survived 0, 1 and 2 years, respectively, since operation. The table provides the posterior mean of this random time quantity, along with 90% credibility intervals, both for the median remaining time after surviving t_0 years and for median total remaining time since operation. See also Fig. 3, which illustrates the relative ease with which one can provide inference for even complicated parameters, operationally (via posterior simulation) and regarding interpretation (the histogram carries more information than a point estimate and a confidence interval, which, incidentally, is a quite difficult task to construct from frequentist large-sample theory).

These illustrations are made with the same a_0 prior guess as above, and with prior sample size $k = 10$. Larger values of k led to larger median residual life length. This is because of the fact that, as noted before, the prior guess underestimates the hazard with respect to what is suggested by the data. That means the stronger the weight of prior guess on posterior samples, the longer the estimated median remaining life length, which corresponds less to the information contained in the data.

8. Bayesian asymptotics: Bernshteĭn–von Mises theorem

In this section, we investigate an asymptotic result for the Bayesian analysis developed in sections 4 and 5. We show that, under the choice of Beta prior for A and minimal conditions for prior $\pi(\gamma)$, the posterior distribution of γ attains asymptotic normality in the form of a

Bernshteĭn–von Mises theorem: $\sqrt{n}(\gamma-\hat{\gamma})\,|\,\text{data}\to_d \mathrm{N}_p(0,\Sigma^{-1})$ in probability, which means that the posterior of γ, centred at the partial likelihood estimator $\hat{\gamma}$, is asymptotically equivalent to the sampling distribution of $\hat{\gamma}-\gamma_{\mathrm{tr}}$.

The Bernshteĭn–von Mises theorem is desirable also from a practical point of view. As far as the Bayesian computational capacity has increased, when traditional methods do not lead to easily implementable algorithms, say for complicated functionals, the Bayesian credible sets represent an alternative to confidence intervals (and vice versa). The Bernshteĭn–von Mises theorem is the theoretical justification of this practice: it implies that Bayesian credible sets reach asymptotically nominal coverage probability like consistent estimation based on maximum partial likelihood.

For parametric models the Bernshteĭn–von Mises theorem is a well-known result (see Schervish, 1995, section 7.4), whereas in non-parametric model it is not automatically guaranteed by posterior consistency of the prior process. For right-censored survival data, the Beta process leads to both consistency and asymptotic normality, as indicated in Hjort (1990b) and later proved in broader generality by Kim & Lee (2001, 2004). Kim & Lee (2003b) proved the Bernshteĭn–von Mises theorem for a Bayesian bootstrap scheme, for the Cox regression model, where the marginal posterior of regression coefficients has the partial likelihood as a likelihood component multiplied with the prior. Bayesian analysis with an informative prior on the baseline hazard leads to a marginal posterior with a fairly complicated integrated likelihood, as shown in section 5. The proof of the Bernshteĭn–von Mises theorem in this setting may follow the arguments of Kim & Lee (2003b, theorem 2), provided that there is high-order asymptotic equivalence of the integrated likelihood with the ordinary partial likelihood. Our proof partly follows these methods.

Recall from theorem 2 that

$$\pi(\gamma\,|\,\text{data})\propto\exp\{\Lambda_n(\gamma)\}\pi(\gamma),\qquad\text{with }\Lambda_n(\gamma)=-\rho_n(\gamma)+\sum_{i=1}^{q_n}\log k_i(\gamma). \tag{34}$$

In the following, we assume that data are generated under model (2) with true parameters $(A_{\mathrm{tr}},\gamma_{\mathrm{tr}})$, as in section 3, and that there are no ties among the complete observations. The posterior density of $h=\sqrt{n}(\gamma-\hat{\gamma})$ is proportional to

$$g_n(h)=\exp\left\{\Lambda_n(\hat{\gamma}+n^{-1/2}h)-\Lambda_n(\hat{\gamma})\right\}\pi(\hat{\gamma}+n^{-1/2}h).$$

It is sufficient to show that $g_n(h)$ converges in L_1-norm to $\psi(h)\pi(\gamma_{\mathrm{tr}})$, where $\psi(h)=\exp(-(1/2)h^{\mathrm{t}}\Sigma h)$. The conditions for the theorem to hold are similar to those of frequentist asymptotics (see De Blasi, 2006).

Theorem 4 (Bernshteĭn–von Mises theorem for γ)
Assume that the regularity conditions underlying the frequentist theory of section 3 are in force, involving in particular a positive definite limit variance matrix Σ*, that the covariates are uniformly bounded, and that the proportion of individuals* $n^{-1}\sum_{i=1}^n Y_i(\tau)$ *alive at the time end-point* τ *stays bounded away from zero in probability. Let also* $\pi(\gamma)$ *be continuous at* γ_{tr} *with* $\pi(\gamma_{\mathrm{tr}})$ *positive. Then*

$$\int_{\mathcal{R}^p}|g_n(h)-\psi(h)\pi(\gamma_{\mathrm{tr}})|\,\mathrm{d}h\to_p 0. \tag{35}$$

The conditions on $\pi(\gamma)$ are standard for Bayesian consistency to hold. It is easily seen that they are fulfilled by the Jeffreys prior of section 4. Bounded covariates ensure that $r(x^{\mathrm{t}}\gamma)$ is bounded away from zero and one when γ runs through a compact set.

The proof of (35) is associated with the following decomposition:

$$\int_{\mathcal{R}^p} \left| g_n(h) - \psi(h)\pi(\gamma_{\rm tr}) \right| \mathrm{d}h \le \int_{|h|\le K} \left| g_n(h) - \psi(h)\pi(\gamma_{\rm tr}) \right| \mathrm{d}h + \int_{K<|h|<\delta\sqrt{n}} g_n(h)\,\mathrm{d}h + \int_{|h|\ge\delta\sqrt{n}} g_n(h)\,\mathrm{d}h + \int_{|h|>K} \psi(h)\pi(\gamma_{\rm tr})\,\mathrm{d}h.$$

As the fourth term can be made arbitrarily small, we need the following three lemmas:

Lemma 5
For any $\epsilon, K>0$, $\Pr\{\int_{|h|\le K} |g_n(h) - \psi(h)\pi(\gamma_{\rm tr})|\,\mathrm{d}h > \epsilon\} \to 0$.

Lemma 6
For any $\epsilon>0$, *there are* $K, \delta>0$ *with* $\Pr\{\int_{K<|h|<\delta\sqrt{n}} g_n(h)\,\mathrm{d}h > \epsilon\} \to 0$.

Lemma 7
For any $\epsilon, \delta>0$, $\Pr\{\int_{|h|\ge\delta\sqrt{n}} g_n(h)\,\mathrm{d}h > \epsilon\} \to 0$.

We provide the proofs of lemmas 5–7 in Appendix A, together with three further lemmas required for showing that $\Lambda_n(\gamma)$ and the log-partial likelihood function are asymptotically equivalent.

In this discussion, we have focussed on the posterior distribution of $\sqrt{n}(\gamma-\hat\gamma)$. To be sure that the Bernshteĭn–von Mises theorem holds for all smooth functionals of (A,γ), like the median residual lifetime in section 7, we need a stronger result, about joint convergence to the correct normal limit of the posterior distribution of the full $\sqrt{n}(A-\hat A, \gamma-\hat\gamma)$. Such a result may also be proved, by an extension of our methods, but space does not allow us to pursue this inside the present article.

9. Concluding remarks

Below we offer some concluding comments, some pointing to further research questions of relevance.

Mixed Cox and logistic relative risk models

Section 2 gave probabilistic and statistical motivation for lifetime distributions with hazard rate of the form $\rho\lambda(s)$, where $\rho \le 1$ depends on shock sizes and $\lambda(s)$ depends on shock frequency. Covariates could influence one or both of these quantities. This motivates more general hazard models of the form

$$\alpha_i(s) = \alpha(s)\exp(u_i^{\rm t}\beta)\frac{\exp(v_i^{\rm t}\gamma)}{1+\exp(v_i^{\rm t}\gamma)},$$

say, where u_i influences the frequency and v_i influences the sizes of shocks. Methods of our article can be generalized to such models.

Time-dependent covariates

Our presentation has focussed on time-independent covariates, as with $\mathrm{d}A_i(s) = \mathrm{d}A(s)r(x_i^{\rm t}\gamma)$, but the framework and our methodology extend without severe difficulties to situations where x_i may depend on time t. Then the cumulative hazard rates are not proportional.

Time-discrete models

Assume lifelength data T_i^0 are recorded only for time points $b, 2b, 3b, \ldots$, say, rather than in a fully time-continuous fashion. We may then work with models for hazard rates of the type $\alpha_i(bs) = \alpha(bs)r(x_i^t\gamma)$ for $s = 1, 2, 3, \ldots$ and $i = 1, \ldots, n$, where $\alpha_i(bs)$ is the probability that $T_i^0 = bs$ given that $T_i^0 \geq bs$. A time-discrete Beta process prior is appropriate, where the $\alpha(bs)$ probabilities are taken independent and Beta distributed. Results can be obtained in analogy to those reached in our article, which can be seen as the limiting case where $b \to 0$.

Neutral to the right priors

We have focussed on Beta processes when it comes to placing priors on the cumulative hazard rate A, but more general Lévy processes can be used as well, like general neutral to the right priors, with machinery for posterior distributions similar to that developed in section 5.

Doubly non-parametric Bayes methods for proportional hazards

A general proportional hazards formulation takes $\mathrm{d}A_i(s) = \mathrm{d}A(s)r_i$ for individuals $i = 1, \ldots, n$. Our emphasis has been on semiparametric models where A is modelled non-parametrically but the r_is are modelled parametrically, say via $r_i = r(x_i^t\gamma)$ for a given r function. The essence of our Bayesian construction is retained with more general ways of modelling $(r_1, \ldots, r_n)$. One may, in particular, investigate non-parametric versions of these too, as with $r_i = G(x_i)$ and a random distribution function G. Such schemes would lead to doubly non-parametric Bayes methods.

Distinguishing between the Cox model and the logistic relative risk model

For our illustration in section 7 we compared Nelson–Aalen plots for $\hat{Z}_i$s and Z_i^*s, computed under models (1) and (2), see (32). One may formalize ways of ascertaining which of the two plots that most closely fit the unit exponential cumulative hazard rate plot, i.e. the diagonal. This would lead to formal statistical ways of distinguishing between the two models. Properties of such tests should then be investigated. One may view these procedures as exposing each of the two models to a goodness-of-fit inspection (see e.g. Hjort, 1990a; McKeague & Utikal, 1991). The $p+1$-dimensional relative risk model (33) offers another way of separating the two models, e.g. via this bigger model's profile log-partial likelihood function, as exemplified in Fig. 1. It is also perfectly possible to analyse data inside this $p+1$-dimensional model. The transformation used is only monotone for $\kappa \leq 1$, so the parameter set for the parameter should most fruitfully be taken as $(-\infty, 1]$ or $[0, 1]$. Yet another angle on the problem is via model selection criteria. Methods of Hjort & Claeskens (2006) may in fact be used to estimate the mean squared error of any parameter estimate, inside each of the two models, after which the model giving smallest risk estimate is preferred.

Competing risks models

Survival data models focus on the aspects of only one important transition, namely from 'alive' to 'dead'. In various situations one needs to distinguish between different outcomes or transitions, as when one records deaths of different types or when accounting for all transitions between 'married', 'unmarried' and 'dead'. A model for such competing risks, where transitions $1, \ldots, k$ need accounting for, can take $A_j(t, x)$ equal to $A(t)\exp(x^t\gamma_j)/\{1 + \exp(x^t\gamma_j)\}$ for $j = 1, \ldots, k$. Such a model can be handled in a semiparametric Bayes fashion, with

methods of this article, using a Beta process for A and parametric priors for the vectors $\gamma_1, \ldots, \gamma_k$. References to situations where such models and approaches may be attempted include Borgan (2002), Fine & Gray (1999) and Andersen *et al.* (2003).

Acknowledgements

The authors are grateful to Ørnulf Borgan and to two referees for constructive comments that led to improvements in the manuscript.

References

Aalen, O. O. & Hjort, N. L. (2002). Frailty models that yield proportional hazards. *Statist. Probab. Lett.* **58**, 335–342.

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–1120.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer, New York.

Andersen, P. K., Klein, J. P. & Rosthøj, S. (2003). Generalized linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90**, 15–27.

Besag, J. & Green, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55**, 25–37.

Borgan, Ø. (2002). Estimation of covariate-dependent Markov transition probabilities from nested case-control data. *Statist. Methods Med. Res.* **11**, 183–202. Correction **12**, 124.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

Damien, P., Laud, P. W. & Smith, A. (1995). Approximate random variate generation from infinitely divisible distribution with application to Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **57**, 547–563.

Damien, P., Laud, P. W. & Smith, A. (1996). Implementation of Bayesian non-parametric inference based on Beta process. *Scand. J. Statist.* **23**, 27–36.

De Blasi, P. (2006). *Semiparametric models in Bayesian event history analysis using Beta processes*. PhD thesis, IMQ, Bocconi University, Milano. Available at: http://www.cees.no/?option=com_staff&person=pierpab.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.

Ferguson, T. S. & Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.* **43**, 1634–1643.

Ferguson, T. S. & Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163–186.

Fine, J. P. & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Amer. Statist. Assoc.* **94**, 496–509.

Gelfand, A. E. & Kottas, A. (2003). Bayesian semiparametric regression for median residual life. *Scand. J. Statist.* **30**, 651–665.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis,* 2nd edn. Chapman & Hall/CRC, London.

Gjessing, H. K., Aalen, O. O. & Hjort, N. L. (2003). Fraily models based on Lévy processes. *Adv. Appl. Probab.* **35**, 532–550.

Gould, S. J. (1995). The median isn't the message. In *Adam's naval and other essays*, 15–21. Penguin Books, New York.

Hjort, N. L. (1986). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scand. J. Statist.* **13**, 63–85.

Hjort, N. L. (1990a). Goodness of fit tests in models for life history data based on cumulative hazard rates. *Ann. Statist.* **18**, 1221–1258.

Hjort, N. L. (1990b). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294.

Hjort, N. L. (2003). Topics in nonparametric Bayesian statistics [with discussion]. In *Highly structured stochastic systems* (eds P. J. Green, N. L. Hjort & S. Richardson), 455–487. Oxford University Press, Oxford.

Hjort, N. L. & Claeskens, G. (2006). Focused information criteria and model averaging for Cox's hazard regression model. *J. Amer. Statist. Assoc.* **101**, 1449–1464.
Hjort, N. L. & Kim, Y. (2007). Beta process and their application to survival and event history analysis. Statistical Research Report, Department of Mathematics, University of Oslo.
Hjort, N. L. & Pollard, D. B. (1993). *Asymptotics for minimisers of convex processes*. Statistical Research Report, Department of Mathematics, University of Oslo, Oslo.
Ishwaran, H. & James, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data. *J. Amer. Statist. Assoc.* **99**, 175–190.
Kalbfleisch, J. D. (1978). Non-parametric analysis of survival time data. *J. Roy. Statist. Soc. Ser. B* **40**, 214–221.
Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Ann. Statist.* **27**, 562–588.
Kim, Y. & Lee, J. (2001). On posterior consistency of survival models. *Ann. Statist.* **29**, 666–686.
Kim, Y. & Lee, J. (2003a). Bayesian analysis of proportional hazard models. *Ann. Statist.* **31**, 493–511.
Kim, Y. & Lee, J. (2003b). Bayesian bootstrap for proportional hazards models. *Ann. Statist.* **31**, 1905–1922.
Kim, Y. & Lee, J. (2004). A Bernstein–von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32**, 1492–1512.
Laud, P. W., Damien, P. & Smith, A. F. M. (1998). Bayesian nonparametric and covariate analysis of failure time data. In *Practical nonparametric and semiparametric Bayesian statistics*, Lecture Notes in Statistics 133. Springer, New York.
Lee, J. & Kim, Y. (2004). A new algorithm to generate Beta processes. *Comput. Statist. Data Anal.* **47**, 441–453.
McKeague, I. W. & Utikal, K. J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards models. *Scand. J. Statist.* **18**, 177–195.
Prentice, R. L. & Self, S. G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Ann. Statist.* **11**, 804–813.
Rigat, F. (2004). *Beta–Stacy survival models and Bayesian Weibull survival trees*. PhD thesis, ISDS, Duke University.
Schervish, M. J. (1995). *Theory of statistics*. Springer, New York.
Walker, S. & Damien, P. (2000). Representation of Levy processes without Gaussian components. *Biometrika* **87**, 477–483.
Walker, S. & Mallick, B. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. Roy. Statist. Soc. Ser. B* **59**, 845–860.
Wild, C. J. & Kalbfleisch, J. D. (1981). A note on a paper of Ferguson and Phadia. *Ann. Statist.* **9**, 1061–1065.
Wolpert, R. L. & Ickstadt, K. (1998). Simulation of Lévy random fields. In *Practical nonparametric and semiparametric Bayesian statistics*, Lecture Notes in Statistics 133, Springer, New York.

Received January 2006, in final form September 2006

Pierpaolo De Blasi, Departimento di Statistica e Matematica Applicata, Università degli Studi di Torino, Piazza Arbarello, 8–10122 Torino, Italy.
E-mail: pierpaolo.deblasi@unito.it

Appendix: proof of lemmas 5 – 7

In the sequel the following notation is needed. For vectors a we write $\|a\| = \sup_i |a_i|$ and $|a| = (a^{\mathrm{t}}a)^{1/2}$, and for matrices A, $\|A\| = \sup_{i,j} |a_{i,j}|$. Let M_x be a constant bounding all covariates $\|x_i\|$, as per theorem 4. Next let

$$S_n^{(0)+}(t,\gamma) = n^{-1} \sum_{j \in R_n^+(t)} r(x_j^{\mathrm{t}}\gamma),$$

$$\ell_n^+(\gamma) = \sum_{i=1}^{q_n} \log \left\{ \frac{r(x_i^{\mathrm{t}}\gamma)}{n S_n^{(0)+}(t_i,\gamma)} \right\}$$

and

$$\Delta_n(\gamma)=\sum_{i=1}^{q_n}\log\left\{\frac{1+r(x_i^{\mathrm{t}}\gamma)}{nS_n^{(0)+}(t_i,\gamma)}\right\},$$

so that $\ell_n^+(\gamma)=\ell_n(\gamma)+\Delta_n(\gamma)$, with $\ell_n(\gamma)$ from (4). Finally $\ell_n^{(k)}$ and $\Lambda_n^{(k)}$ denote the arrays of the kth derivatives of $\ell_n(\gamma)$ and $\Lambda_n(\gamma)$, the latter from (34).

Lemma 8

For any compact subset $\mathcal{G}$ of $\mathcal{R}^p$, and for $k=0,1,2,3$,

$$\sup_{\gamma\in\mathcal{G}}\|\Lambda_n^{(k)}(\gamma)-\ell_n^{(k)}(\gamma)\|=O_p(\log n).$$

Proof. Note that

$$\Lambda_n(\gamma)-\ell_n(\gamma)=-\rho_n(\gamma)+\Delta_n(\gamma)+\sum_{i=1}^{q_n}\log\{1+\xi_i(\gamma)\},$$

with

$$\xi_i(\gamma)=S_n^{(0)+}(t_i,\gamma)\left\{n\int_0^1 R_{n,t_i}^+(s,\gamma)(1-s)^{c(t_i)-1}\,\mathrm{d}s-S_n^{(0)+}(t_i,\gamma)^{-1}\right\}. \tag{36}$$

For $r_*=\inf_{\gamma\in\mathcal{G},\,\|x\|\le M_x} r(x^{\mathrm{t}}\gamma)>0$, we can bound $\sup_{\gamma\in\mathcal{G}}|\Delta_n(\gamma)|$ by $\sum_{i=1}^{q_n}\log[1+\{r_*\sum_{j=1}^n\times Y_j(\tau)\}^{-1}]$, which is $O_p(1)$, as $\sum_{j\in R_n^+(t_i)}1\ge\sum_{j=1}^n Y_j(\tau)$ and $\sum_{j=1}^n Y_j(\tau)/n$ stays away from zero. Similar arguments show that $\sup_{\gamma\in\mathcal{G}}\|\Delta_n^{(k)}(\gamma)\|=O_p(1)$ for $k=1,2,3$. Lemma 10 (which we prove below) implies that $\sup_{\gamma\in\mathcal{G},\,1\le i\le q_n}\|(\partial^k/\partial\gamma^k)\log(1+\xi_i(\gamma))\|$ is $O_p(n^{-1})$ for $k=0,1,2,3$, so that lemma 9 (also to be proven below) completes the proof.

Lemma 9

For any compact subset $\mathcal{G}$ of $\mathcal{R}^p$, and for $k=0,1,2,3$,

$$\sup_{\gamma\in\mathcal{G}}\|\rho_n^{(k)}(\gamma)\|=O_p(\log n).$$

Proof. Let $c_*=\inf_{t\in[0,\,\tau]}c(t)>0$, then

$$g^*=\sup_{t\in[0,\,\tau],\,s\in[0,\,1]}(1-s)^{1-c_*}c(t)(1-s)^{c(t)-1}<\infty.$$

For $r_*=\inf_{\gamma\in\mathcal{G},\,\|x\|\le M_x} r(x^{\mathrm{t}}\gamma)$, we have that

$$\prod_{j=i+1}^{n}(1-r_j s)\le \mathrm{e}^{-(n-i)r_* s},$$

where $r_j=r(x_j^{\mathrm{t}}\gamma)$. For $c_*<1$, $r_*>0$, $\epsilon<1$ and $k=0,1,2$ it can be shown that the sequence of function $g_n(u)=I\{u<n\}\,\mathrm{e}^{-r_* u}u^k(1-u/n)^{c-1}$ is uniformly integrable (see De Blasi 2006, Appendix).

For $k=0$ note that

$$\rho_n(\gamma)\le g^*A_0(\tau)\sum_{i=1}^{n}\int_0^1 \mathrm{e}^{-(n-i)r_* s}(1-s)^{c_*-1}\,\mathrm{d}s\le g^*A_0(\tau)\sum_{i=1}^{n-1}\frac{a_i}{i},$$

where

$$a_i = \int_0^\infty I\{u \le i\}\, \mathrm{e}^{-r_* u} \left(1 - \frac{u}{i}\right)^{c_*-1} \mathrm{d}u.$$

Next $a_i \to 1/r_*$, as the sequence of integrands converges pointwise to $I\{u<\infty\}\,\mathrm{e}^{-r_* u}$ and is uniformly integrable. Hence $g^* A_0(\tau)\sum_{i=1}^{n-1} a_i = O(\log n)$.

For $k=1$, differentiation leads to

$$\rho_n^{(1)}(\gamma) = \sum_{i=1}^n \int_0^{T_i}\int_0^1 x_i r_i(1-r_i) \prod_{j=i+1}^n (1-r_j s) a(s,t)\,\mathrm{d}s\,\mathrm{d}A_0(t) \tag{37}$$

$$+ \sum_{i=1}^n \int_0^{T_i}\int_0^1 r_i \frac{\partial}{\partial\gamma} \prod_{j=i+1}^n (1-r_j s) a(s,t)\,\mathrm{d}s\,\mathrm{d}A_0(t). \tag{38}$$

The term (37) can be treated similarly to the $k=0$ case, as $\|x_i r_i(1-r_i)\| \le M_x/4$. Next, $\|(38)\|$ is dominated by

$$g^* A_0(\tau) \sum_{i=1}^n \int_0^1 \mathrm{e}^{-(n-i)r_* s} \sum_{j=i+1}^n \left\{ \|x_i\| r_i(1-r_i)\left(s + \frac{s^2}{1-r_j}\right)\right\} (1-s)^{c_*-1}\,\mathrm{d}s.$$

For $r^* = \sup_{\gamma\in\mathscr{G},\,\|x\|\le M_x} r(x^{\mathrm{t}}\gamma) < 1$,

$$\sup_{\gamma\in\mathscr{G}} \|(38)\| \le (1/4) M_x g^* A_0(\tau) \sum_{i=1}^{n-1} \frac{b_i}{i},$$

with

$$b_i = \int_0^\infty I\{u \le i\}\, \mathrm{e}^{-r_* u} \left\{ u + \frac{u^2}{i(1-r^*)}\right\} \left(1 - \frac{u}{i}\right)^{c_*-1} \mathrm{d}u.$$

As $1/(1-r^*) < \infty$, the sequence of integrands within $\{b_i\}$ converges pointwise to $I\{u<\infty\} u\,\mathrm{e}^{-r_* u}$, and is uniformly integrable, so $b_i \to 1/r_*^2$. Thus $b_i = O(1)$ and the proof is complete for $k=1$. The results for $k=2,3$ follow via similar arguments.

Lemma 10

For any compact subset $\mathscr{G}$ of $\mathcal{R}^p$, and for $k=0,1,2,3$,

$$\sup_{\gamma\in\mathscr{G},\,1\le i\le q_n} \|\xi_i^{(k)}(\gamma)\| = O_p(n^{-1}). \tag{39}$$

Proof. Write $\xi_i(\gamma) = S_n^{(0)+}(t_i,\gamma)\zeta_i(\gamma)$ where

$$\zeta_i(\gamma) = n\int_0^1 R^+_{n,t_i}(s,\gamma)(1-s)^{c(t_i)-1}\,\mathrm{d}s - S_n^{(0)+}(t_i,\gamma)^{-1}.$$

Note that $|\xi_i(\gamma)| \le |\zeta_i(\gamma)|$ because $\sup_{\gamma\in\mathscr{G},\,1\le i\le q_n} |S_n^{(0)+}(t_i,\gamma)| \le 1$.

We consider the case $k=0$ first. Let $\{r_{n,j}\}$ be the triangular array such that $r_{n,j} = r(x_j^{\mathrm{t}}\gamma)$ for $j \in R_n^+(t_i)$, so that the sequence of row-means corresponds to $\bar r_n = S_n^{(0)+}(t_i,\gamma)$. Note that $\{r_n\}_{n\ge1}$ is bounded away from zero for $\gamma\in\mathscr{G}$ and, for each $t\in[0,\tau]$, $\sum_{R_n^+(t)} 1 = O_p(n)$. Hence, in order to prove that $\sup_{\gamma\in\mathscr{G},\,1\le i\le q_n} |\zeta_i(\gamma)| = O_p(n^{-1})$, it suffices to show that

$$\left| n \int_0^1 \prod_{j=1}^{n} (1 - r_j s)(1-s)^{c-1} \, \mathrm{d}s - \frac{1}{\bar{r}_n} \right| = O(n^{-1}) \tag{40}$$

for any $c > 0$, see De Blasi (2006, Appendix).

For $k = 1$ note that

$$\begin{aligned} \sup_{\gamma \in \mathcal{G},\, 1 \le i \le q_n} \|\xi_i^{(1)}(\gamma)| &\le \sup_{\gamma \in \mathcal{G},\, 1 \le i \le q_n} \|\zeta_i^{(1)}(\gamma) S_n^{+(0)}(t_i, \gamma) - \zeta_i(\gamma) S_n^{+(1)}(t_i, \gamma)\| \\ &\le \sup_{\gamma \in \mathcal{G},\, 1 \le i \le q_n} \|\zeta_i^{(1)}(\gamma)\| + \frac{M_x}{4} \sup_{\gamma \in \mathcal{G},\, 1 \le i \le q_n} |\zeta_i(\gamma)| \end{aligned} \tag{41}$$

for $S_n^{+(1)}(t_i, \gamma)$ defined in analogy to $S_n^{+(0)}(t_i, \gamma)$ and the notation in section 3, while

$$\zeta_i^{(1)}(\gamma) = n \int_0^1 \frac{\partial}{\partial \gamma} R^+_{n,\, t_i}(s, \gamma)(1-s)^{c(t_i)-1} \, \mathrm{d}s + S_n^{+(1)}(t_i, \gamma) S_n^{+(0)}(t_i, \gamma)^{-2}.$$

We may treat the second term of (41) as in the $k = 0$ case, while for the first term note that $\partial/\partial\gamma R^+_{n,\, t_i}(s, \gamma)$ can be expressed as $-R^+_{n,\, t_i}(s, \gamma) n S_n^{(1)+}(t_i, \gamma)\{s + O(s^2)\}$, for $0 < O(s^2) \le s^2/(1 - r^*)$ (see the definition of r^* above), so that

$$\zeta_i^{(1)}(\gamma) = -S_n^{(1)+}(t_i, \gamma) \left[n \int_0^n R^+_{n,\, t_i}(s, \gamma) n \{s + O(s^2)\}(1-s)^{c(t_i)-1} \, \mathrm{d}s - S_n^{+(0)}(t_i, \gamma)^{-2} \right].$$

As $\sup_{t \in [0,\, \tau],\, \gamma \in \mathcal{G}} \|S_n^{(1)+}(t, \gamma)\| = O_p(1)$, for the term in square brackets we may apply a result similar to (40) for the array $\{r_{n,j}\}$, namely that

$$\left| n \int_0^1 \prod_{j=1}^{n} (1 - r_j s) n \left\{ s + O(s^2) \right\} (1-s)^{c-1} \, \mathrm{d}s - \frac{1}{\bar{r}_n^2} \right| = O(n^{-1}), \tag{42}$$

for any $c > 0$ and such that the $O(s^2)$ term in question is bounded by some M (see De Blasi, 2006, Appendix). Similar arguments can be applied for $k = 2, 3$.

Proof of lemma 5

Consider the Taylor expansion $\Lambda_n(\gamma)$ around $\hat{\gamma}$:

$$\Lambda_n(\hat{\gamma} + n^{-1/2} h) - \Lambda_n(\hat{\gamma}) = h^{\mathrm{t}} \frac{\Lambda_n^{(1)}(\hat{\gamma})}{\sqrt{n}} - \frac{1}{2} h^{\mathrm{t}} \left(-\frac{1}{n} \Lambda_n^{(2)}(\hat{\gamma}) \right) h + Q_n(h, \gamma_*), \tag{43}$$

for

$$|\gamma_* - \hat{\gamma}| \le \frac{h}{\sqrt{n}} \quad \text{and} \quad Q_n(h, \gamma) = \sum_{j,\, l,\, k} h_j h_k h_l \Lambda^{(3)}_{n,j,l,k}(\gamma) \frac{1}{6 n^{3/2}}.$$

Note that, for $\psi_n(h) = \exp[-(1/2) h^{\mathrm{t}} \{-\Lambda_n^{(2)}(\hat{\gamma})/n\}]$,

$$\begin{aligned} \int_{|h| \le K} |g_n(h) - \psi(h)\pi(\gamma_{\mathrm{tr}})| \, \mathrm{d}h &\le \pi(\gamma_{\mathrm{tr}}) \int_{|h| \le K} |\psi_n(h) - \psi(h)| \, \mathrm{d}h \\ &\quad + \int_{|h| \le K} |g_n(h) - \psi_n(h)\pi(\gamma_{\mathrm{tr}})| \, \mathrm{d}h. \end{aligned} \tag{44}$$

Lemma 8 and consistency of the observed information matrix $-n^{-1} \ell_n^{(2)}(\hat{\gamma})$ (see De Blasi, 2006) imply that $\|\Lambda_n^{(2)}(\hat{\gamma})/n + \Sigma\| \to_p 0$, which in concert with $\pi(\gamma_{\mathrm{tr}}) > 0$ implies that $\sup_{|h| \le K} \pi(\gamma_{\mathrm{tr}}) |\psi_n(h) - \psi(h)| \to_p 0$. For the second term, on $|h| \le K$,

$$|g_n(h) - \psi_n(h)\pi(\gamma_{\mathrm{tr}})| \le \psi_n(h)\pi(\gamma_{\mathrm{tr}})[(1 + \eta_1)\eta_2 + \eta_1],$$

where

$$\eta_1 = \sup_{|h|\le K} \left|\pi(\hat\gamma + n^{-1/2}h)/\pi(\gamma_{\rm tr}) - 1\right|, \quad \eta_2 = \sup_{|h|\le K} \left|\exp\left(\frac{h^{\rm t}\Lambda_n^{(1)}(\hat\gamma)}{\sqrt{n}} + Q_n(h,\gamma_*)\right) - 1\right|.$$

Note that $\eta_1 \to_p 0$ because of $\pi\left(\hat\gamma + h/\sqrt{n}\right) \to_p \pi(\gamma_{\rm tr})$ uniformly on $|h| \le K$ and the hypothesis on $\pi(\gamma_{\rm tr})$. In order to show that also $\eta_2 \to_p 0$, consider first that $\sup_{|h|\le K} |h^{\rm t}\Lambda_n^{(1)}(\hat\gamma)/\sqrt{n}| \le K\|\Lambda_n^{(1)}(\hat\gamma)/\sqrt{n}\|$ and lemma 8 implies that $\|\Lambda_n^{(1)}(\hat\gamma)\| = o_p(\sqrt{n})$. Secondly, $\sup_{|h|\le K} |\, Q_n(h,\gamma_*)\,| \le \|\Lambda_n^{(3)}(\gamma_*)/n\|p^3K^3/6\sqrt{n}$ and lemma 8 implies that $\|\Lambda_n^{(3)}(\gamma_*)\| = O_p(n)$, as long as $n^{-1}\ell_n^{(3)}(\gamma,\tau)$ has a well-defined limit in probability in compact neighbourhood of $\gamma_{\rm tr}$, which can be proved with standard martingale arguments under the assumptions taken. Finally, $\|\Lambda_n^{(2)}(\hat\gamma)/n + \Sigma\| \to_p 0$ and the positive definitiveness of Σ implies that (44) converges to zero in probability.

Proof of lemma 6

For any δ, $|h| < \delta\sqrt{n}$ implies that $|\gamma - \gamma_{\rm tr}| \le 2\delta$ eventually, so that there exists $M < \infty$ such that $\Pr\{\|\Lambda_n^{(3)}(\gamma_*)/n\| > M\} \to 0$ (see the proof of lemma 5). With $\kappa_n = n^{-1/2}\Lambda_n^{(1)}(\hat\gamma)$ and $B_n = -\Lambda_n^{(2)}(\hat\gamma)/n - \Sigma$, consider that (a) on $|h| > 1$, $h^{\rm t}\kappa_n \le \|\kappa_n\|h^{\rm t}h$; (b) for λ the smallest eigenvalue of Σ, $h^{\rm t}\Sigma h \ge \lambda h^{\rm t}h$, while $|h^{\rm t}B_nh| \le p^2\,\|B_n\|\,h^{\rm t}h$; (c) on $\|\Lambda_n^{(3)}(\gamma_*)/n\| \le M$, $|Q_n(h,\gamma_*)| \le p^3\delta M h^{\rm t}h/6$. Hence, on $\|\Lambda_n^{(3)}(\gamma_*)/n\| \le M$, expansion in (43) leads to the following bound:

$$\Lambda_n(\hat\gamma + n^{-1/2}h) - \Lambda_n(\hat\gamma) \le -\tfrac{1}{2}(-2\|\kappa_n\| + \lambda - p^2\|B_n\| - p^3\delta M/3)h^{\rm t}h.$$

Next note that $\lambda > 0$ because Σ is positive definite and that $\|\kappa_n\| \to_p 0$ together with $\|B_n\| \to_p 0$ (see the proof of lemma 5) guarantees that, for small ϵ_1, there exists $\delta = \delta(\lambda, M, \epsilon_1)$ such that $\lambda - p^3\delta M/3 - \epsilon_1 > 0$ and

$$\Pr\{-2\|\kappa_n\| + \lambda - p^2\|B_n\| - p^3\delta M/3 < \epsilon_1\} \to 0. \tag{45}$$

For given $\epsilon > 0$, $\Pr\{\int_{K<|h|<\delta\sqrt{n}} g_n(h)\,{\rm d}h > \epsilon\}$ can be bounded by

$$\Pr\left\{\sup_{|h|\le\delta\sqrt{n}} \left|\pi(\hat\gamma + h/\sqrt{n})/\pi(\gamma_{\rm tr})\right| > \rho\right\} + \Pr\{\|\Lambda_n^{(3)}(\gamma_*)/n\| > M\}$$
$$+ \Pr\{p^2\|B_n\| + \kappa_n/2 > (\lambda - p^3\delta M/3 - \epsilon_1)\} + \Pr\left\{\int_{K<|h|<\delta\sqrt{n}} {\rm e}^{-\epsilon_1 h^{\rm t}h/2}\,{\rm d}h > \epsilon/\rho\pi(\gamma_{\rm tr})\right\},$$

where $\rho = \sup_{|h|\le 2\delta} |\,\pi(\gamma_{\rm tr} + h)/\pi(\gamma_{\rm tr})|$. We choose $K > 1$ such that $\int_{|h|>K} {\rm e}^{-\epsilon_1 h^{\rm t}h/2}\,{\rm d}h \le \epsilon/\rho\pi(\gamma_{\rm tr})$. Finally (45) and $\Pr\left\{\sup_{|h|\le\delta\sqrt{n}} \left|\pi(\hat\gamma + h/\sqrt{n})/\pi(\gamma_{\rm tr})\right| > \rho\right\} \to 0$ imply that there exist δ such that the first three terms go to zero, and this ends the proof.

Proof of lemma 7

As $\int_{|h|\ge\delta\sqrt{n}} g_n(h)\,{\rm d}h \le n^{p/2}\sup_{|\gamma-\hat\gamma|\ge\delta} \exp\{\Lambda_n(\gamma) - \Lambda_n(\hat\gamma)\}$ and $|\gamma - \hat\gamma| \ge \delta$ implies that $|\gamma - \gamma_{\rm tr}| \ge \delta/2$ eventually, it suffices to prove that, for a given constant m, that we determine later,

$$n^{p/2} \sup_{\delta/2\le|\gamma-\gamma_{\rm tr}|\le m} \exp\{\Lambda_n(\gamma) - \Lambda_n(\hat\gamma)\} = o_p(1), \tag{46}$$

$$n^{p/2} \sup_{|\gamma-\gamma_{\rm tr}|>m} \exp\{\Lambda_n(\gamma) - \Lambda_n(\hat\gamma)\} = o_p(1). \tag{47}$$

For (46), we exploit the fact that $n^{-1}\{\ell_n(\gamma) - \ell_n(\gamma_{\rm tr})\}$ converges uniformly on compact sets to $d(\gamma)$, which is a negative function with a unique maximum at $\gamma_{\rm tr}$ such that $d(\gamma_{\rm tr}) = 0$ (see De Blasi, 2006). Two applications of lemma 8 and the consistency of $\hat\gamma$ lead to

$$n^{p/2} \sup_{\delta/2\le|\gamma-\gamma_{\rm tr}|\le m} \exp\{\Lambda_n(\gamma) - \Lambda_n(\hat\gamma)\} \le n^{p/2} \sup_{\delta/2\le|\gamma-\gamma_{\rm tr}|\le m} \exp\{n[o_p(1) + d(\gamma)]\},$$

which shows that (46) holds, as $n^{p/2}\exp\{nd(\gamma)\} \to 0$.

For (47), as $n^{-1}\{\Lambda_n(\gamma)-\Lambda_n(\hat{\gamma})\}$ is not concave, we construct a sequence of concave functions that play upper bounds and whose tail areas vanish eventually in probability. Note that under the condition of covariates centred around zero, it can be shown that $\ell_n^+(\gamma)$ is asymptotically a concave function. We have that

$$\begin{aligned}\Lambda_n(\gamma) &\le \ell_n^+(\gamma)+\sum_{i=1}^{q_n}\log\left[nS_n^{(0)+}(t_i,\gamma)\int_0^1 R^+_{n,t_i}(s,\gamma)(1-s)^{c(t_i)-1}\,\mathrm{d}s\right]\\ &\le \ell_n^+(\gamma)+\sum_{i=1}^{q_n}\log\left[S_n^{(0)+}(t_i,\gamma)\int_0^n \mathrm{e}^{-S_n^{(0)+}(t_i,\gamma)u}\left(\frac{1-u}{n}\right)^{c(t_i)-1}\mathrm{d}u\right],\end{aligned}$$

where the last inequality holds as $\prod_{j\in R_n^+(t)}\{1-r(x_j^{\mathrm{t}}\gamma)s\}$ is bounded by $\exp\{-\sum_{j\in R_n^+(t)} r(x_j^{\mathrm{t}}\gamma)s\}=\exp\{-nS_n^{(0)+}(t,\gamma)s\}$ and via a change of variable. Furthermore,

$$\begin{aligned}&\sum_{i=1}^{q_n}\log\left[S_n^{(0)+}(t_i,\gamma)\int_0^n \mathrm{e}^{-S_n^{(0)+}(t_i,\gamma)u}\left(1-\frac{u}{n}\right)^{c(t_i)-1}\mathrm{d}u\right]\\ &\quad\le\sum_{i=1}^{q_n}\log\left[1+S_n^{(0)+}(t_i,\gamma)\int_0^n \mathrm{e}^{-S_n^{(0)+}(t_i,\gamma)u}\left\{\left(1-\frac{u}{n}\right)^{c(t_i)-1}-1\right\}\mathrm{d}u\right]\le q_n\log(1+K_n^*),\end{aligned}$$

where

$$K_n^*=\max_{i=1,\ldots,q_n,\,\gamma,\,n}\left[0,S_n^{(0)+}(t_i,\gamma)\int_0^n \mathrm{e}^{-S_n^{(0)+}(t_i,\gamma)u}\left\{\left(1-\frac{u}{n}\right)^{c(t_i)-1}-1\right\}\mathrm{d}u\right].$$

We next show that K_n^* stays bounded in n. As $K_n^*=0$ for $c(t_i)\ge 1$, it suffices to consider integrals like $r_n\int_0^n \mathrm{e}^{-r_nu}\{(1-u/n)^{c-1}-1\}\,\mathrm{d}u$ for any $0<c<1$ and for any sequence r_n of real numbers in $[0,1]$. For each s, the sequence $nr_n\mathrm{e}^{-nr_ns}$ attains its maximum at e^{-1}/s for $nr_n=1/s$. Finally, uniformly in n,

$$r_n\int_0^n \mathrm{e}^{-r_nu}\left\{\left(\frac{1-u}{n}\right)^{c-1}-1\right\}\mathrm{d}u\le \mathrm{e}^{-1}\int_0^1 s^{-1}\{(1-s)^{c-1}-1\}\,\mathrm{d}s<\infty.$$

Next $\ell_n^+(\gamma)=\ell_n(\gamma)+\Delta_n(\gamma)$ with $\sup_{\gamma\in\mathscr{G}}|\Delta_n(\gamma)|=O_p(1)$, see the proof of lemma 8, and $n^{-1}\{\ell_n^+(\gamma)-\ell_n^+(\gamma_{\mathrm{tr}})\}$ converges to $d(\gamma)$ uniformly on compact set. Finally, note that $q_n/n\to_p q=\Pr\{N_1(\tau)=1\}$ and we may choose m such that, for a real $\eta>0$, $\sup_{|\gamma-\gamma_{\mathrm{tr}}|=m} d(\gamma)\le -qK_n^*-\eta$. By Lemma 8, we arrive at

$$\begin{aligned}\sup_{|\gamma-\gamma_{\mathrm{tr}}|>m}\mathrm{e}^{\Lambda_n(\gamma)-\Lambda_n(\hat{\gamma})}&\le\sup_{|\gamma-\gamma_{\mathrm{tr}}|>m}\exp\left[n\{q_nK_n^*/n+n^{-1}\ell_n^+(\gamma)-n^{-1}\Lambda_n(\hat{\gamma})\}\right]\\ &\le\exp\left[n\{qK_n^*+\sup_{|\gamma-\gamma_{\mathrm{tr}}|=m}d(\gamma)+o_p(1)\}\right]\le\exp\left[n\{-\eta+o_p(1)\}\right].\end{aligned}$$

The proof of (47) is completed by noting that $n^{p/2}\exp(-n\eta)\to 0$.